STA6800: Week01 - Chapter 01 Introduction

Ick Hoon Jin

2021-08-10

Why Networks?

- **Network Everywhere**
  - Statement that "we live in a connected world" captures why networks have come to hold such interest in recent years.
  - From on-line social networks like Facebook to the World Wide Web and the Internet itself, we are surrounded by examples of ways in which we interact with each other.
  - Similarly, we are connected as well at the level of various human institutions (e.g., governments), processes (e.g., economies), and infrastructures (e.g., the global airline network).
  - The image of a network is a natural one to use to capture the notion of elements in a system and their interconnectedness.
- **Network Definition**
  - The term 'network' seems to be used in a variety of ways, at various levels of formality.
  - The Oxford English Dictionary defines the word network in its most general form simply as "a collection of interconnected things."
  - 'network' is used inter-changeably with the term 'graph' since, for mathematical purposes, networks are most commonly represented in a formal manner using graphs of various kinds.
- **Graph**
  - Graph: A structure amounting to a set of objects in which some pairs of the objects are in some sense "related".
  - Vertices: The objects correspond to mathematical abstractions.
  - Edges: Each of the related pairs of vertices.
  - The edges may be directed or undirected.
    - If the vertices represent people at a party, and there is an edge between two people if they shake hands, then this graph is undirected because any person A can shake hands with a

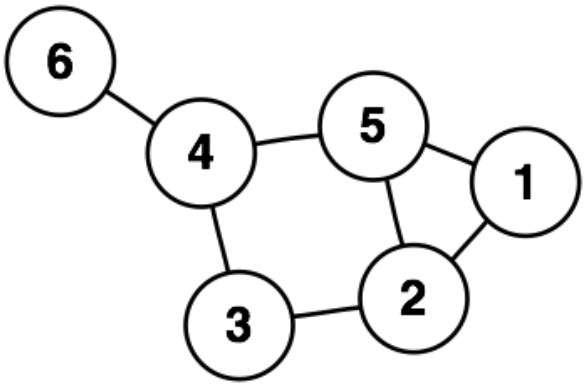

Figure 1: A Graph with Six Vertices and Seven Edges

person B only if B also shakes hands with A.

* In contrast, if any edge from a person A to a person B corresponds to A owes money to B, then this graph is directed, because owing money is not necessarily reciprocated.

– The former type of graph is called an undirected graph while the latter type of graph is called a directed graph.

- The seeds of network-based analysis: The 1735 solution of Euler to the famous Konigsberg bridge problem, in which he proved that it was impossible to walk the seven bridges of that city in such a way as to traverse each only once.

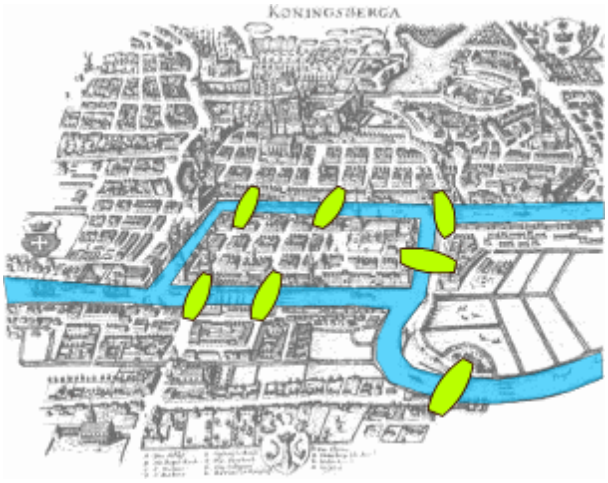

Figure 2: Map of Konigsberg in Euler's time showing the actual layout of the seven bridges, highlighting the river Pregel and the bridges

- Applications

  – The theory of electrical circuits in electric engineering.

  – The study of molecular structure in chemistry.

  – Transportation, allocation in the fields of operations research and computer science.

  - Characterizing interactions within social groups in sociology.
  - Modeling and analysis of complex systems: Statistical physics and computer science.
    * Statistical physics can be attributed a seminal role in encouraging what has now become a pervasive emphasis across the sciences on understanding how the interacting behaviors of constituent parts of a whole system lead to collective behavior and systems-level properties or outcomes.
    * Computer science can be attributed much of the theory and methodology for conceptualizing, storing, manipulating, and doing computations with networks and related data, particularly in ways that enable efficient handling of the often massive quantities of such data.
      · Information networks (e.g, the World Wide Web)
      · Related social media applications (e.g., Twitter)
  - A network-based perspective recently has been found to be useful in the study of complex systems across a diverse range of application areas.
    * Computational biology (e.g., studying systems of interacting genes, proteins, chemical compounds, or organisms)
    * Engineering (e.g., establishing how best to design and deploy a network of sensing devices)
    * Finance (e.g., studying the interplay among, say, the world's banks as part of the global economy),
    * Marketing (e.g., assessing the extent to which product adoption can be induced as a type of 'contagion'),
    * Neuroscience (e.g., exploring patterns of voltage dynamics in the brain associated with epileptic seizures)
    * political science (e.g., studying how voting preferences in a group evolve in the face of various internal and external forces),
    * and public health (e.g., studying the spread of infectious disease in a population, and how best to control that spread)

- Two important contributing factors to the phenomenal growth of interest in networks
  1. an increasing tendency towards a systems-level perspective in the sciences, away from the reductionism that characterized much of the previous century, and
  2. an accompanying facility for high-throughput data collection, storage, and management.

- Object this course

- The study of systems such as those just described is accompanied by measurement and, accordingly, the need for statistical analysis.
- The focus of this course is on how to use tools in R to do statistical analysis of network data.
- More specifically, we aim to present tools for performing what are arguably a core set of analyses of measurements that are either of or from a system conceptualized as a network.

Types of Network Analysis

Visualizing and Characterizing Networks

- Visualization
  - The visualization and numerical characterization of a network usually is one of the first steps in network analysis.
  - Lazega's Lawyer Network

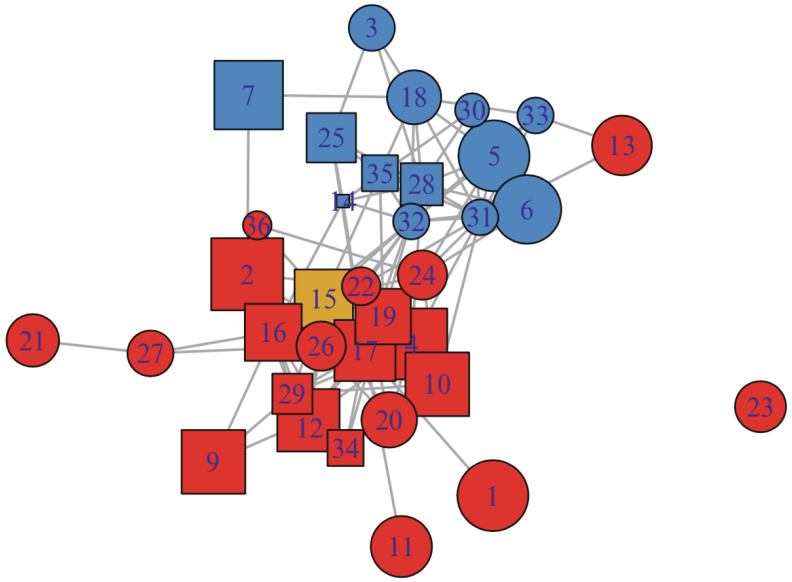

Fig. 1.1 Visualization of Lazega's network of collaborative working relationships among lawyers. *Vertices* represent partners and are labeled according to their seniority (with 1 being most senior). Vertex colors (i.e., *red*, *blue*, and *yellow*) indicate three different office locations, while vertex shape corresponds to the type of practice [i.e., litigation (*circle*) and corporate (*square*)]. Vertex area is proportional to number of years with the law firm. *Edges* indicate collaboration between partners. There are three female partners (i.e., those with seniority labels 27, 29, and 34); the rest are male

    * Visualization of part of a dataset on collaborative working relationships among members of a New England law firm.
    * These data were collected for the purpose of studying cooperation among social actors in an organization, through the exchange of various types of resources among them.

    - The organization observed was a law firm, consisting of over 70 lawyers (roughly half partners and the other half associates) in three offices located in three different cities.
    - Relational data reflecting resource exchange were collected, and additional attribute information was recorded for each lawyer, including type of practice, gender, and seniority.
  - Visual Summary of Lazega's Lawyer Network
    - A graph is used to represent the network, with vertices corresponding to lawyers, and edges, to collaboration between pairs of lawyers.
    - In addition, differences in vertex color, shape, size, and label are used to indicate office location, type of practice, years with the law firm, and seniority.
  - The visualization of large networks is a separate challenge of its own.
- Numerical Summaries
  - Characterization of network data through numerical summaries is another important aspect of descriptive analysis for networks.
  - Summary measures for networks necessarily seek to capture characteristics of a graph.
  - Transitivity (Clustering Coefficient)
    - It is natural to ask to what extent two lawyers that both work with a third lawyer are likely to work with each other as well.
    - This notion corresponds to the social network concept of transitivity and can be captured numerically through an enumeration of the proportion of vertex triples that form triangles (i.e., all three vertex pairs are connected by edges), typically summarized in a so-called clustering coefficient.
  - Assortativity Coefficient
    - Given the two types of lawyers represented in these data (i.e., corporate and litigation), it is natural to ask to what extent lawyers of each type collaborate with those of same versus different types.
    - This notion corresponds to the social network concept of assortativity and can be quantified by a type of correlation statistic (the so-called assortativity coefficient), in which labels of connected pairs of vertices are compared.
    - The focus is on an attribute associated with network vertices (i.e., lawyer practice) and the network structure plays a comparatively more implicit role.
- The ranges from characterization of properties associated with individual vertices or edges to properties of subgraphs to properties of the graph as a whole.

Network Modeling and Inference

- Beyond asking what an observed network looks like and characterizing its structure, at a more fundamental level we may be interested in understanding how it may have arisen. That is, we can conceive of the network as having resulted from some underlying processes associated with the complex system of interest to us and ask what are the essential aspects of these processes.
- The actual manner in which the network was obtained, i.e., the corresponding measurement and construction process, may well be important to take into consideration.
- Network Modeling: There are two classes of network models: mathematical and statistical network models.
  - Mathematical Models: Models specified through (typically) simple probabilistic rules for generating a network, where often the rules are defined in an attempt to capture a particular mechanism or principle (e.g., 'the rich get richer').
  - Statistical Models: Models (often probabilistic as well) specified at least in part with the intention that they be fit to observed data and that the fit be effected and assessed using formal principles of statistical inference.
  - While certainly there is some overlap between these two classes of models, the relevant literatures nevertheless are largely distinct.
- Erdos-Renyi Model
  - Model in which edges are assigned randomly to pairs of vertices based on the result of a collection of independent and identically distributed coin tosses—one toss for each vertex pair.
  - Corresponding to a variant of the famous Erdos-Renyi formulation of a random graph, this model has been studied extensively since the 1960s.
  - Its strength is in the fact not only that its properties are so well understood (e.g., in terms of how, for example, cohesive structure emerges as a function of the probability of an edge) but also in the role it plays as a standard against which to compare other, more complicated models.
- Mathematical Network Model
  - Mathematical models generally are too simple to be a good match to real network data.
  - Nevertheless, they are not only useful in allowing for formal insight to be gained into how specific mechanisms of edge formation may affect network structure, but they also are used commonly in defining null classes of networks against which to assess the 'significance' of structural characteristics found in an observed network.
- Statistical Network Models

  - Exponential random graph models are analogous to generalized linear models, being based on an exponential family form.
  - Latent network models, in specifying that edges may arise at least in part from an unmeasured (and possibly unknown) variable(s), directly parallel the use of latent variables in hierarchical modeling.
  - Stochastic block models may be viewed as a form of mixture model.
  - Nevertheless, importantly, the specification of such models and their fitting typically are decidedly less standard, given the usually high-dimensional and dependent nature of the data.

Network Processes

- Network Processes
  - As objects for representing interactions among elements of a complex system, network graphs are frequently the primary focus of network analysis.
  - It is actually some quantity (or attribute) associated with each of the elements in the system that ultimately is of most interest.
  - Nevertheless, in such settings it often is not unreasonable to expect that this quantity be influenced in an important manner by the interactions among the elements, and hence the network graph may still be relevant for modeling and analysis.
  - We can picture a stochastic process as 'living' on the network and indexed by the vertices in the network.
  - A variety of questions regarding such processes can be interpreted as problems of prediction of either static or dynamic network processes.
- Dynamic Processes
  - Many of the systems studied from a network-based perspective are intrinsically dynamic in nature.
  - Many processes defined on networks are more accurately thought of as dynamic, rather than static, processes.
  - Example: The context of public health and disease control.
    * Understanding the spread of a disease (e.g., the H1N1 flu virus) through a population can be modeled as the diffusion of a binary dynamic process (indicating infected or not infected) through a network graph, in which vertices represent individuals, and edges, contact between individuals.
  - Mathematical modeling (using both deterministic and stochastic models) arguably is still the primary tool for modeling such processes, but network-based statistical models gradually

are seeing increased use, particularly as better and more extensive data on contact networks becomes available.

- Statistical methods for analyzing network flows. Referring to the movement of something — materials, people, or commodities, for example — from origin to destination, flows are a special type of dynamic process fundamental to transportation networks (e.g., airlines moving people) and communication networks (e.g., the Internet moving packets of information), among others.
- Dynamic network analysis, wherein the network, the process(es) on the network, or indeed both, are expected to be evolving in time.

Talking about Graphs

Basic Graph Concepts

- *A Simple Graph*: A graph has no edges for which both ends connect to a single vertex (called *loops*) and no pairs of vertices with more than one edge between them (called *multi-edges*). Its edges are referred to as proper edges.
- An object with either of these properties is called a *multi-graph*.
- The most basic notion of connectivity is that of adjacency. Two vertices $u, v \in V$ are said to be adjacent if joined by an edge in E. Such vertices are also referred to as *neighbors*.
- Two edges $e_1, e_2 \in E$ are adjacent if joined by a common endpoint in V. A vertex $v \in V$ is incident on an edge $e \in E$ if v is an end point of e. From this follows the notion of the degree of a vertex v, say d_v, defined as the number of edges incident on v.
- For digraphs, vertex degree is replaced by in-degree (i.e., d_v^{in}) and out-degree (i.e., d_v^{out}), which count the number of edges pointing in towards and out from a vertex, respectively.

The concept of movement about a graph.

- A *walk* on a graph G, from v_0 to v_l, is an alternating sequence $\{v_0, e_1, v_1, e_2, \cdots, v_{l-1}, e_l, v_l\}$, where the endpoints of e_i are $\{v_{i-1}, v_i\}$.
- The length of this walk is said to be l.
- Refinements of a walk include *trails*, which are walks without repeated edges, and *paths*, which are trails without repeated vertices.
- A trail for which the beginning and ending vertices are the same is called a *circuit*.
- A walk of length at least three, for which the beginning and ending vertices are the same, but for which all other vertices are distinct from each other, is called a *cycle*.
- Graphs containing no cycles are called *acyclic*.

- A vertex v in a graph G is said to be *reachable* from another vertex u if there exists a walk from u to v.
- The graph G is said to be *connected* if every vertex is reachable from every other. A component of a graph is a maximally connected subgraph. That is, it is a connected subgraph of G for which the addition of any other remaining vertex in V would ruin the property of connectivity.
- A digraph G is weakly connected if its underlying graph (i.e., the result of stripping away the labels tail and head from G) is connected. It is called strongly connected if every vertex v is reachable from every u by a directed walk.
- A common notion of *distance* between vertices on a graph is defined as the length of the shortest path(s) between the vertices (which we set equal to infinity if no such path exists). This distance is often referred to as *geodesic distance*, with 'geodesic' being another name for shortest paths. The value of the longest distance in a graph is called the *diameter* of the graph.

Special Types of Graphs

- A *complete* graph is a graph where every vertex is joined to every other vertex by an edge. This concept is perhaps most useful in practice through its role in defining a clique, which is a complete subgraph.

```
library(igraph)
```

```
##
## Attaching package: 'igraph'
## The following objects are masked from 'package:stats':
##
##     decompose, spectrum
## The following object is masked from 'package:base':
##
##     union
```

```
g.full <- graph.full(7)
plot(g.full)
```

- A *regular* graph is a graph in which every vertex has the same degree. A regular graph with common degree d is called *d-regular*.

```
g.ring <- graph.ring(7)
plot(g.ring)
```

- A connected graph with no cycles is called a *tree*. The disjoint union of such graphs is called a *forest*.

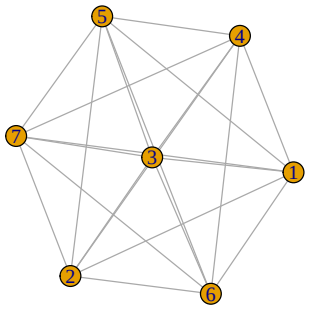

Figure 3: An Example of the Complete Graph

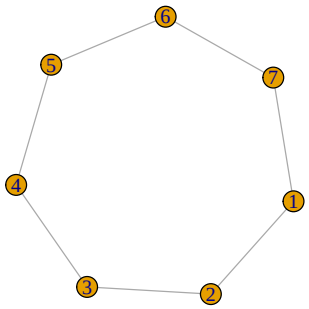

Figure 4: An Example of the Regular Graph

```
g.tree <- graph.tree(7, children=2, mode="undirected")
plot(g.tree)
```

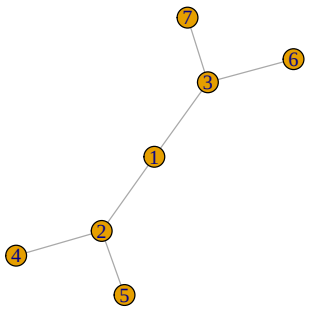

Figure 5: An Example of the Tree Graph

- Trees are of fundamental importance in the analysis of networks. They serve, for example, as a key data structure in the efficient design of many computational algorithms.
  - A digraph whose underlying graph is a tree is called a *directed tree*.
  - Often such trees have associated with them a special vertex called a *root*, which is distinguished by being the only vertex from which there is a directed path to every other vertex in the graph. Such a graph is called a *rooted tree*.
  - A vertex preceding another vertex on a path from the root is called an *ancestor*, while a vertex following another vertex is called a *descendant*. Immediate ancestors are called *parents*, and immediate descendants, *children*. A vertex without any children is called a *leaf*. The distance from the root to the farthest leaf is called the *depth* of the tree.

- A k-star is a special case of a tree, consisting only of one root and k leaves.

```
g.star <- graph.star(7, mode="undirected")
plot(g.star)
```

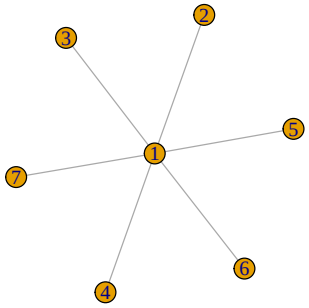

Figure 6: An Example of the Star Graph

Such graphs are useful for conceptualizing a vertex and its immediate neighbors (ignoring any connectivity among the neighbors).

An important generalization of the concept of a tree is that of a directed acyclic graph (i.e., the DAG). A DAG, as its name implies, is a graph that is directed and that has no directed cycles. However, unlike a directed tree, its underlying graph is not necessarily a tree, in that replacing the arcs with undirected edges may leave a graph that contains cycles.

A bipartite graph is a graph $G = (V, E)$ such that the vertex set V may be partitioned into two disjoint sets, say V_1 and V_2, and each edge in E has one endpoint in V_1 and the other in V_2.

Such graphs typically are used to represent `membership' networks, for example, with`members' denoted by vertices in V_1, and the corresponding 'organizations', by vertices in V_2.

```
g.bip = graph.formula(actor1:actor2:actor3, movie1:movie2,
                      actor1:actor2 - movie1,
                      actor2:actor3 - movie2)
V(g.bip)$type = grepl("^movie", V(g.bip)$name)
print_all(g.bip, v=T)
```

```
## IGRAPH 1cf2cee UN-B 5 4 --
## + attr: name (v/c), type (v/l)
## + vertex attributes:
## |       name  type
## | [1] actor1 FALSE
## | [2] actor2 FALSE
## | [3] actor3 FALSE
## | [4] movie1  TRUE
## | [5] movie2  TRUE
## + edges from 1cf2cee (vertex names):
```

```
## [1] actor1--movie1 actor2--movie1 actor2--movie2 actor3--movie2
```

It is not uncommon to accompany a bipartite graph with at least one of two possible induced graphs. Specifically, a graph $G_1 = (V_1, E_1)$ may be defined on the vertex set V_1 by assigning an edge to any pair of vertices that both have edges in E to at least one common vertex in V_2. Similarly, a graph G_2 may be defined on V_2. Each of these graphs is called a projection onto its corresponding vertex subset.

```
proj <- bipartite.projection(g.bip)
print_all(proj[[1]])
```

```
## IGRAPH 40aad01 UNW- 3 2 --
## + attr: name (v/c), weight (e/n)
## + edges from 40aad01 (vertex names):
## [1] actor1--actor2 actor2--actor3
```

```
print_all(proj[[2]])
```

```
## IGRAPH 645d4d5 UNW- 2 1 --
## + attr: name (v/c), weight (e/n)
## + edge from 645d4d5 (vertex names):
## [1] movie1--movie2
```