Analyzing Ecological Networks

EcoNet Group

https://econetoolbox.github.io/

Cesab - April 2024

Part I

Introduction to networks

Outline Part 1

Examples of graphs I

Figure: A social network.

Examples of graphs II

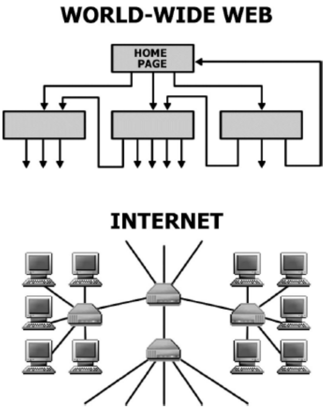


Figure: Internet and WWW. Source: [1].

Examples of graphs III

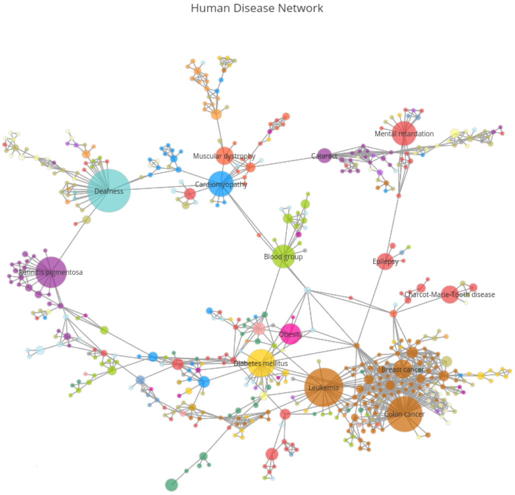


Figure: Gene regulatory network of human diseases.

Examples of graphs IV

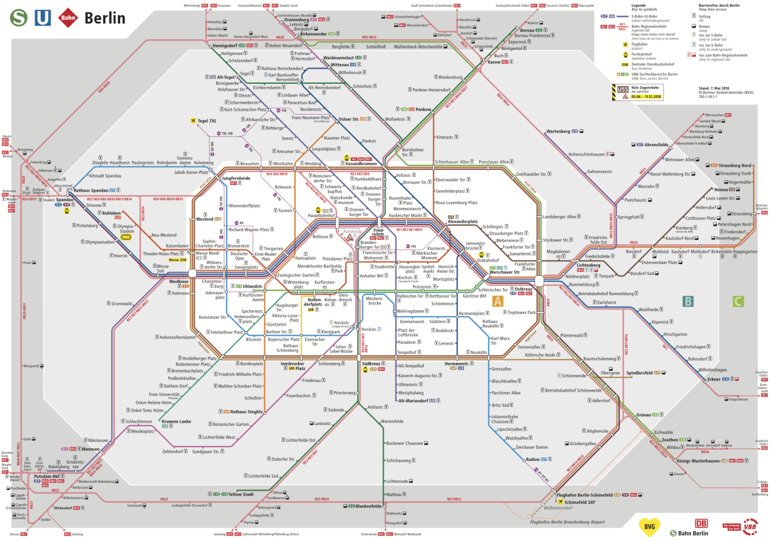


Figure: Subway network in Berlin.

Examples of graphs (foll.) I

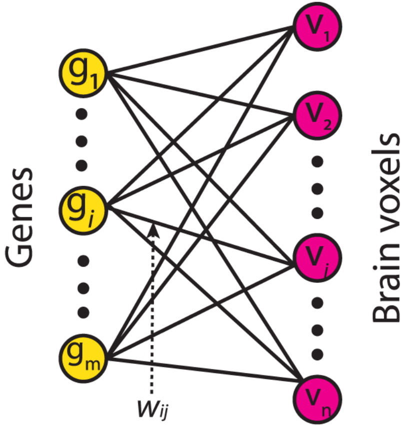


Figure: Bipartite networks of genes and brain voxels. Source: [3].

Examples of graphs (foll.) II

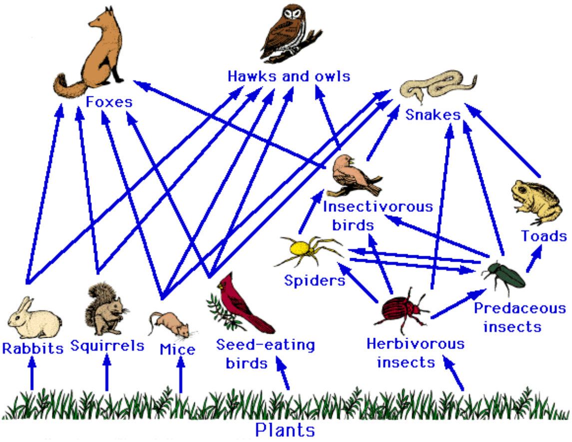


Figure: Simplified trophic network (food web). A directed link indicates who is the prey of whom.

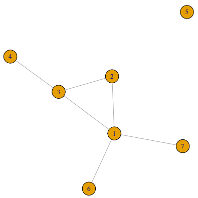
Vocabulary - Basic definitions

- A graph $G = (V, E)$ is a set of nodes (or vertices) $V = \{1, \ldots, n\}$ and a set of edges (or links) $E \subset V^2$
- n is the order; $|E|$ is the size
- graphs can be undirected ($\{i, j\} \in E$) or directed ($(i, j) \in E$); binary (edge $\{i, j\}$ is present or absent) or weighted (present edge $\{i, j\}$ has a value w_{ij}; when $w_{ij} \in \mathbb{N}$ this is a multiplicity); with or without self-loops ($\{i, i\}$ is a self-loop);
- a node is isolated if it doesn't belong to any edge;
- a bipartite graph is s.t. $V = V_1 \cup V_2$ and $V_1 \cap V_2 = \emptyset$ and edges $e = \{u, v\} \in E$ are such that $u \in V_1, v \in V_2$ (e.g. bipartite network of genes and brain voxels)

Data structures

- Adjacency matrix $A = (A_{ij})_{i,j \in V}$ where $A_{ij} = 1\{\{i,j\} \in E\}$ (or $A_{ij} = w_{ij}$)
  - Undirected graphs have symmetric adjacency matrices
  - when graphs are sparse (ie not too many edges), this representation as a matrix is not efficient (n^2 size);
- List of edges: this encoding is the most efficient.
  - NB: if the list of nodes is not additionally given, there cannot be isolated nodes;

$$\begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$



1,2
1,3
1,6
1,7
2,3
3,4

Data structures - Bipartite case

- A bipartite graph has $n_T = n_1 + n_2$ nodes. Its adjacency matrix A is $n_T \times n_T$ with zero block diagonals

$$\left(\begin{array}{c|c} \mathbf{0} & \tilde{\mathbf{A}} \\ \hline \tilde{\mathbf{A}}^\intercal & \mathbf{0} \end{array}\right)$$

- In ecology, the matrix $\tilde{A}$ of size $n_1 \times n_2$ is called incidence matrix.
- Warning: in maths & CS terminology, the incidence matrix H is a $|V| \times |E|$ matrix with entries $H_{ie} = 1$ when node $i \in V$ belongs to edge $e \in E$, and 0 otherwise.

Outline Part 1

Different visualisations of the same graph I

Warning: Visualisation can be misleading!

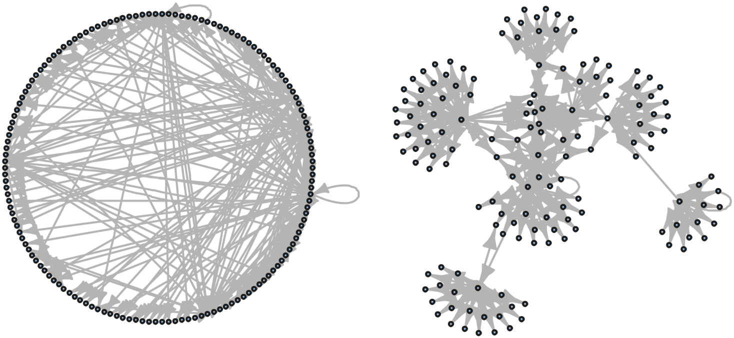


Figure: 2 representations of the same blogs network [4].

Different visualisations of the same graph II

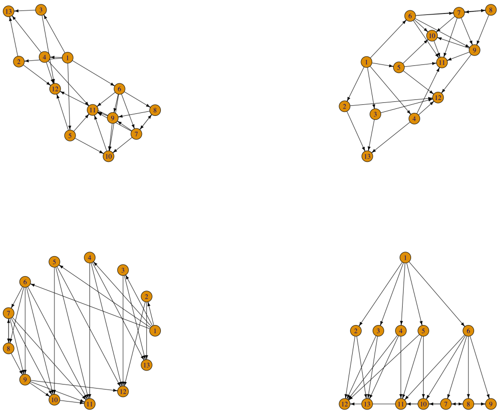


Figure: Different visualisations of the food web from Figure 6.

Different visualisations of the same graph III

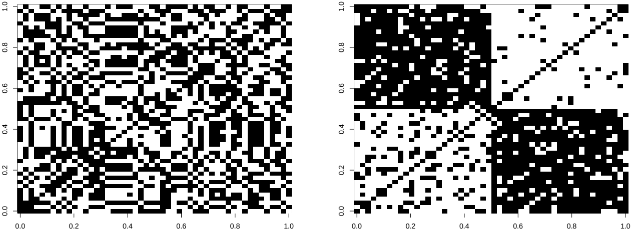
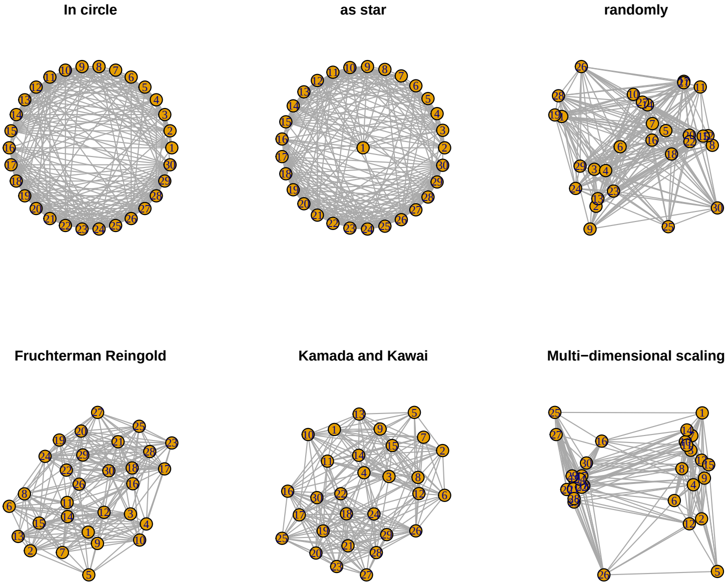


Figure: Dotplot representation of a graph: random node numbering (left) and specific permutation of the nodes (right)

Examples of representations



Outline Part 1

Density / Connectance

A simple binary graph has at most $\binom{n}{2} = n(n-1)/2$ edges.
Its density or connectance is:

$$\text{den}(G) = \frac{|E|}{\binom{n}{2}} = \frac{|E|}{n(n-1)/2}.$$

- the complete graph K_n is the undirected graph with n nodes that contains all possible $\binom{n}{2}$ edges; it has density 1.
- a clique is a complete subgraph in a graph

Neighbors and degrees I

- Neighbors of node $i \in V$ are $\mathcal{N}_i = \{j \in V, j \neq i, \{i, j\} \in E\}$: nodes connected to i in the graph
- Degree of node i is the number of its neighbours $d_i = |\mathcal{N}_i| = \sum_{j \neq i} A_{ij} = \sum_{j \neq i} A_{ji}$
- In directed graphs, one may define indegrees and outdegrees: $d_i^{out} = \sum_{j \neq i} A_{ij}$ and $d_i^{in} = \sum_{j \neq i} A_{ji}$
- Degrees are obtained as rowSums or colSums of adjacency matrix
- We always have $\sum_{i=1}^n d_i = 2|E|$
- Average degree $\bar{d} = n^{-1} \sum_{i=1}^n d_i$
- a d-regular graph has constant degree d (ex infinite grid)
- Hubs (informal) a hub is a large degree node in a graph

Neighbors and degrees II

Degree distributions only loosely characterize graphs

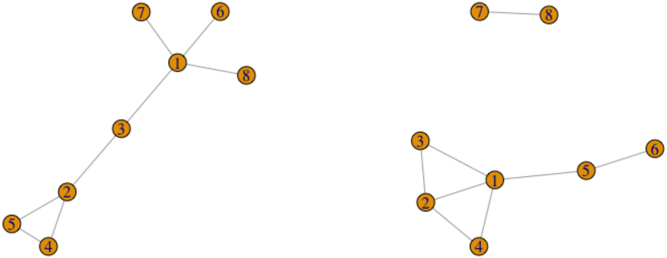
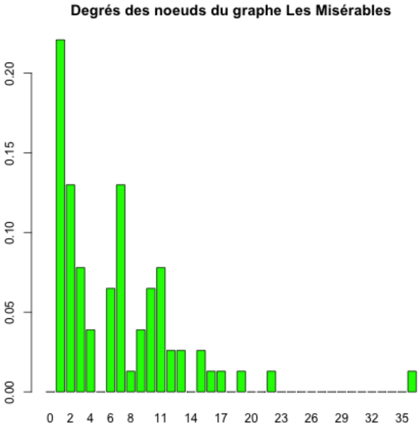


Figure: Example of 2 graphs with same degree sequence.

Neighbors and degrees III

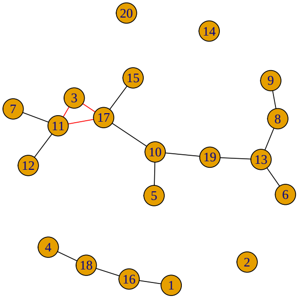
Graphs often show degree distributions with heavy tails, such as scale-free distributions



Paths, connectivity, diameter I

Paths

- A path between nodes $i, j \in V$ is a sequence of edges $e_1, \ldots, e_k \in E$ such that e_t and e_{t+1} share a node, $i \in e_1$ and $j \in e_k$. Its length is k;
- A cycle is a path that connects a node to itself; (ex: a self-loop is a cycle of length 1)



Paths, connectivity, diameter II

Connectivity

- A set of nodes $C = \{v_1, \ldots, v_k\} \in V$ such that there exists a path between any 2 nodes $v_i, v_j \in C$ is a connected component (cc);
- Any graph may be decomposed into a unique collection of maximal cc;
- An isolated node forms a (maximal) cc;
- There are at most $n - |E|$ such maximal cc;
- When there is a unique cc, the graph is connected;
- Giant component (informal): In a sequence of graphs G_n each with n nodes, let C_n be the largest mcc in G_n. We say that C_n is a giant component if its relative size $|C_n|/n$ does not tend to 0 as n increases;

Paths, connectivity, diameter III

Diameter

- the distance ℓ_{ij} between 2 nodes $i, j \in V$ is the length of the shortest path between i, j (and $+\infty$ if the nodes are not in the same cc)
- the average distance in the graph is $\bar{\ell} = 1/(n(n-1)) \sum_{i,j} \ell_{ij}$
- diameter $\mathrm{diam}(G) = \max\{\ell_{ij}; i, j \in V\}$;
- It's finite only if the graph is connected;
- Small-world property (informal): a graph has the small-world property whenever $\bar{\ell}$ is of the order of $\log(n)$;
- See the small-world experiment by Stanley Milgram; and its modern version: three and a half degrees of separation [2]

Clustering coefficients, transitivity, centrality I

Friends of my friends are my friends ...

- Clustering coefficient C_i is the number of edges $|E_i|$ between neighbors of node i divided by the maximum of such number $d_i(d_i-1)/2$; *i.e.*

$$C_i = \begin{cases} \frac{2|E_i|}{d_i(d_i-1)} & \text{if } d_i \geq 2, \\ 0 & \text{otherwise} \end{cases}$$

- It is the connectance of the subgraph induced by the neighbors of i; thus $C_i \in [0,1]$
- the average clustering coefficient is $\bar{C} = \frac{1}{|V|}\sum_{i\in V} C_i$
- Transitivity is

$$T = \frac{\text{Nb of triangles}}{\text{Nb of triplets of connected nodes}}$$

Clustering coefficients, transitivity, centrality II

Friends of my friends are my friends ...

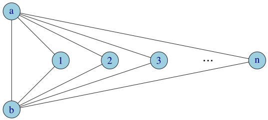


Figure: Here $C_i = 1$ for all nodes except a, b and thus $\bar{C}$ tends to 1. However T tends to 0.

Clustering coefficients, transitivity, centrality III

Friends of my friends are my friends . . .

Centrality

- Degree centrality $C_D(i) = d_i$
- Closeness centrality $C_P(i) = \left(\sum_{j \in V} \ell_{ij}\right)^{-1}$, where ℓ_{ij} is the distance between i, j
- Betweenness centrality $C_B(i) = \sum_{j,k:j \neq k \neq i} \frac{g_{jk}(i)}{g_{jk}}$, where g_{jk} is the number of shortest paths from j to k, and $g_{jk}(i)$ is the number of shortest paths from j to k that go through i;

Motifs I

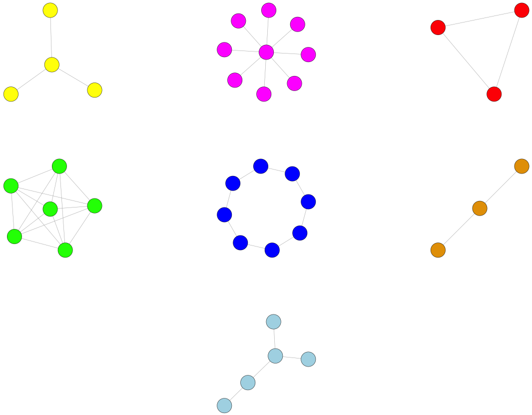


Figure: Examples of motifs: stars (k-stars with $k = 3$ and $k = 8$), cliques (K_3 or triangle and K_6), cycle of length 8, ...

Motifs II

- Counting frequencies of small sizes motifs may be a way to characterize the topology of the graph;
- When the size of the motif becomes large, enumerating all occurrences of a motif becomes a computationally difficult problem;
- with a null model, one can test the hypothesis that the observed frequencies of a motif are too large or too small wrt to some expected value;

Outline Part 1

Outline Part 1

How can we sample interaction data?

Motivations

- Your dataset is sampled in some way from a more complex system;
- Sampling data interactions can be done in many ways, leading to various bias;
- Necessary to understand the sampling scheme and thus the potential bias!

2 fundamental questions

- How my interactions dataset has been sampled and which bias does this create?
- Do the characteristics of my dataset represent the characteristics of the larger unobserved complex system? (difficult question, no general answer)

What impact? Let's take an example

- We are interested in expected degree $\mathbb{E}(D)$.
- $G = (V, E)$ is the observed graph, sampled from an unknown and larger $G^\star = (V^\star, E^\star)$ with $|V^\star| = n^\star$ nodes.
- Sampling scheme: Assume nodes from G are taken uniformly among those in $G^\star$ and for each sampled node $i \in V$, either
  - 1st case: you can observe the interactions $(i, j) \in E^\star$ even if j has not been sampled, i.e. $j \notin V$
  - 2nd case: you observe the interactions $(i, j) \in E$ only if both $i, j \in V$ and $(i, j) \in E^\star$,
- In the second case, the degree $d_i^{(2)} << d_i^{(1)}$.
- average degrees $\bar{D}^{(1)} = \frac{1}{n}\sum_{i=1}^{n} d_i^{(1)}$ and $\bar{D}^{(2)} = \frac{1}{n}\sum_{i=1}^{n} d_i^{(2)}$ are in general very different and $\bar{D}^{(2)}$ underestimates $\mathbb{E}(D)$.

Outline Part 1

Induced and incident subgraphs samplings I

Induced subgraph sampling

- Sample n individuals without replacement from the existing $n^\star$ nodes and observe the links between these nodes.
- Example: you select a set of species and you record all known trophic interactions between them.
- Remarks:
  - you did not select the species uniformly among all possible ones (otherwise, most probably, the graph would be empty)
  - you might (or not) be able to estimate the probability of sampling each individual ($n^\star$ might be unknown);
  - there might be interactions that are unknown from you (additional error in observing the interaction, once the nodes are sampled).

Induced and incident subgraphs samplings II

Incident subgraph sampling

- Sample m edges without replacement from the existing $m^\star$ edges, with each node incident to an edge being included in the graph.
- Example: you have a database of recorded interactions (trophic, mutualistic, ...) and you sample interactions in that database. Or you observe interactions (on the field) among all existing ones;
- Remarks:
  - there is no isolated node in that graph;
  - you might (or not) be able to estimate the probability of sampling each interaction;
  - you should observe in general a low average degree because few edges are incident to the same nodes.

Link tracing sampling schemes I

General principle: Sample n individuals without replacement from the existing $n^\star$ nodes and follow paths from these nodes.

Egonetwork

- Observe edges incident to the initial set of nodes (paths of length 1)
- 2 variants: either include or not the neighbor nodes in the graph.
- Example: You select some plant species and observe their interactions with pollinators. You might identify or not the pollinator (in general, you do).
- Remarks:
  - egonetworks might look like a collection of stars;
  - In theory, you observe all interactions of the selected nodes so the observed degree is the true degree.

Link tracing sampling schemes II

Snowball sampling

- Iterated egonetwork sampling: start with V_0 nodes and observe incident edges. Incident nodes are denoted V_1, then observe edges incident to $V_1 \cup V_0$. New incident nodes are called V_2, etc ....
- Stop either when V_k is empty (all actors have been sampled), or after K iterations.
- Final graph has $V = V_0 \cup V_1 \cup \cdots \cup V_K$ nodes and its edges are either all or a subset of the edges from true graph $G^\star$ that are incident to nodes in V.
- Examples: Web crawling; examples in ecology? ...
- Remarks: Important degree bias: after the first step, it's more likely that you recruit a node with large degree.

Conclusions on sampling schemes

- It is important to select a sampling scheme that is adapted to the type of data AND to the questions explored.
- Keep in mind that your observed statistics might be biased due to the sampling scheme (most of the time, difficult to correct for that)

References I

[1] Albert, R. and A.-L. Barabási.
Statistical mechanics of complex networks.
Rev. Mod. Phys. 74, 47–97, 2002.

[2] Bhagat, S., M. Burke, C. Diuk, I. O. Filiz, and S. Edunov (2016).
Three and a half degrees of separation.
facebook research blog https://research.fb.com/three-and-a-half-degrees-of-separation/.

[3] Ji, S., W. Zhang, and R. Li
A probabilistic latent semantic analysis model for coclustering the mouse brain atlas.
IEEE/ACM Transactions on Computational Biology and Bioinformatics 10(6), 1460–1468, 2014.

References II

[4] Kolaczyk, E. D. and G. Csárdi (2014).
Statistical analysis of network data with R.
Use R! Springer, New York.

[5] V. Krebs.
Unloaking terrorist networks.
Connections, 24(3), 2001.