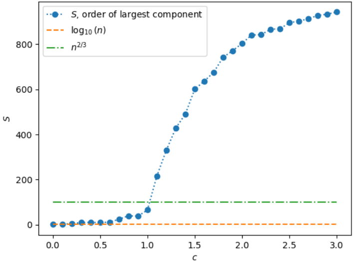
Week 10, Part 1: Giant Components and Small Worlds

Dr Niall Madden

School of Maths, University of Galway

19+20 March 2025)



Homework Assignment 2

Homework Assignment 2 has started

- **Part 1**: A written (i.e., Python-free) assignment. See https://www.niallmadden.ie/2425-CS4423/#Assignment-2-1
- **Part 2**: See https://www.niallmadden.ie/2425-CS4423/#Assignment-2-2
- **Deadline:** 5pm. Friday, 28 March.

Questions?

Outline

This weeks notes are split between PDF slides, and a Jupyter Notebook.

Slides are at:
https://www.niallmadden.ie/2425-CS4423

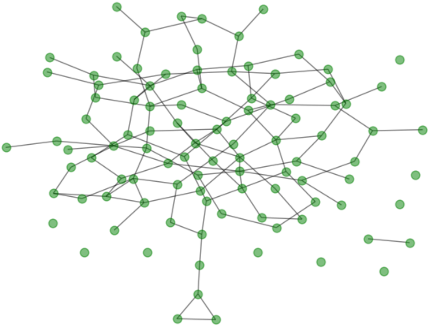
Giant Components

Recall that a network may be made up of several **connected components**, and any connected network has a single connected components.

Giant Components

It is common in large networks to observe a **giant component**: a connected component which has a large proportion of the networks nodes. This is particularly the case with graphs in $G_{ER}(n, p)$ with large enough p. In the following examples we take $n = 100$.

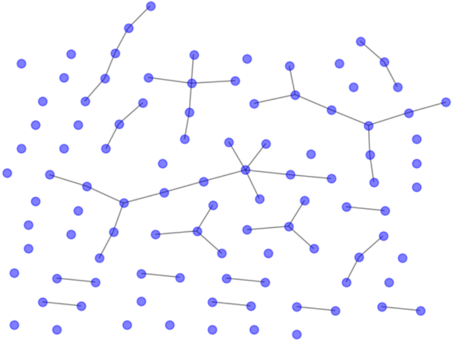
$p = 2/n$; largest component has 89 nodes



Giant Components

It is common in large networks to observe a **giant component**: a connected component which has a large proportion of the networks nodes. This is particularly the case with graphs in $G_{ER}(n, p)$ with large enough p. In the following examples we take $n = 100$.

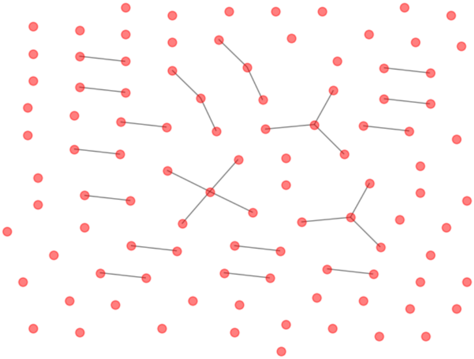
$p = 1/n$; largest component has 13 nodes



Giant Components

It is common in large networks to observe a **giant component**: a connected component which has a large proportion of the networks nodes. This is particularly the case with graphs in $G_{ER}(n, p)$ with large enough p. In the following examples we take $n = 100$.

$p = 0.5/n$; largest component has 5 nodes



Giant component

A connected component of a graph G is called a **giant component** if its number of nodes increases with the order n of G as some positive power of n.

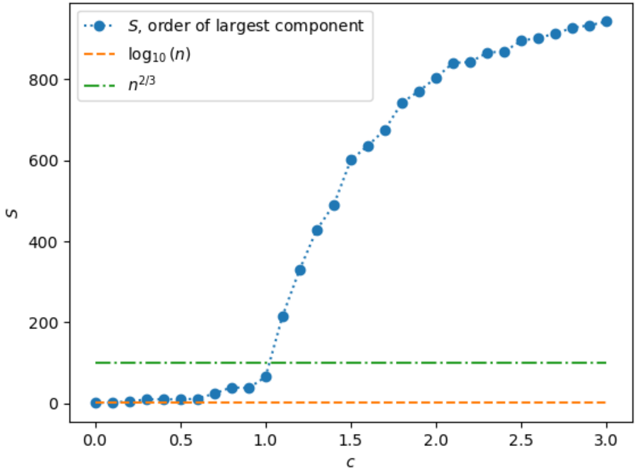
Suppose $p(n) = cn^{-1}$ for some positive constant c. (Then the average degree $\langle k \rangle = pn = c$ remains fixed as $n \to \infty$.)

Theorem (Erdős-Rényi)

For graphs in $G_{ER}(n,p)$:

- *If $c < 1$ the graph contains many small components, orders bounded by $O(\ln n)$.*
- *$c = 1$ the graph has large components of order $S = O(n^{2/3})$.*
- *$c > 1$ there's a unique* **giant component** *of order $S = O(n)$.*

$n = 1000,\ p = cn^{-1}$


S, order of largest component
log10 (n)
n2/3
0
200
400
600
800
S
0.0
0.5
1.0
1.5
2.0
2.5
3.0
c

Small world network

Many real world networks are **small world networks**, where most pairs of nodes are only a few steps away from each other, and where nodes tend to form cliques, i.e., subgraphs having all nodes connected to each other.

Examples:

- MathSciNet allows users to explore distances between authors in the collaborations network. The distance of an author to Erdös is know as this author's Erdös number
- The cinematographic version of this phenomenon is the Six Degrees of Kevin Bacon

Paul Erdös was a prolific mathematics, with over 1,500 published papers, and a prolific collaborator, with over 500 collaborators. The concept of an **Erdös Number** was invented to celebrate the his propensity for collaboration.

Paul Erdös and Terry Tao

- Erdös Number 0: you are Paul Erdös;
- Erdös Number 1: you co-authored a paper with Paul Erdös;
- Erdös Number 2: you co-authored a paper with someone with Erdös Number 1 (and you are not Paul Erdös);
- More generally, your Erdös Number is 1 plus the minimum Erdös Number of your co-authors.

The point of the exercise is to show how **connected** the mathematical world is. E.g., my own EN is 4; the median EN of my colleagues in Mathematics here in Galway is, I believe, 3.

Measures

Three network attributes that measure these small-world effects

- **characteristic path length**, L, defined as the average length of all shortest paths in the network;
- **transitivity**, T, defined as the proportion of *triads* that form triangles;
- **clustering coefficient** C, defined as the *average node clustering coefficient*

Measures

Small worlds networks

A network is called a **small world network** if it has

1. a small *average shortest path length*, L (scaling with $\log n$, where n is the number of nodes), and
2. a high *clustering coefficient*, C.

It turns out that ER random networks do have a small average shortest path length, but not a high clustering coefficient. This observation justifies the need for a different model of random networks, if they are to be used to model the clustering behavior of real world networks.

Distance

We have seen how BFS can determine the length of a shortest path from a given node x to any node y in a *connected network*. An application to all nodes x yields the shortest distances between all pairs of nodes.

Recall (from Week 7, Part 1) that the **distance matrix** of a connected graph $G = (X, E)$, is $\mathcal{D} = (d_{ij})$ where entry d_{ij} is the length of the shortest path from node $i \in X$ to node $j \in X$. (Note: $d_{ii} = 0$ for all i.)

There are a number of graph (and node) attributes that can be defined in terms of this matrix.

Eccentricity: e_i of a node $i \in X$ is the maximum distance between i and any other vertex in G. So, $e_i = \max_j d_{ij}$.

Graph Radius: R is the minimum eccentricity: $R = \min_i e_i$.

Graph Diameter: D is the maximum eccentricity:

$$D = \max_i e_i = -\max_{ij} d_{ij}$$

Note: don't think in terms of "diameter is twice the radius", but rather:

- Diameter is the distance between the points furthest from each other;
- Radius is the distance from the "centre" to the furthest point from it.
- Can be helpful to think about P_n.

Example

The (m, n)-lolipop graph is made from K_n connected to P_n. Sketch the $(3, 3)$-lolipop graph. Write down the distance matrix for this graph. Compute the eccentricity of each node, and then the graph radius and diameter.

Characteristic path length

Definition (Characteristic path length)

The **characteristic path length**, (a.k.a., *average shortest path length*) L, of G is the average distance between pairs of nodes:

$$L = \frac{1}{n(n-1)} \sum_i \sum_j d_{ij}.$$

In tomorrow's class, we'll look at computing the characteristic path length in practice, and in particular for graphs drawn from $G_{ER}(n, m)$ and $G_{ER}(n, p)$.

Spoiler! For these models, $L = \dfrac{\ln n}{\ln\langle k\rangle}$.

Clustering

(As mentioned in Assignment 2, Part 2) In contrast to random graphs, real world networks also contain **many triangles**: it is not uncommon that a friend of one of my friends is my friend, too. This **degree of transitivity** can be measured in several different ways.

For the first we need two concepts:

- The number of **triangles** in G, denoted n_{Δ}, is the number of subgraphs of G that are isomorphic to C_3.
- The number of **triads** in G, denoted $n_{\wedge}$, is the number of pairs of edges with a shared node.

There is an easy way to count the number of **triads** in a network:

- If node i has degree $k_i = \deg(i)$, then it is involved in $\binom{k_i}{2}$ triads;
- So, the total number of triads is $n_\wedge = \sum_i \binom{k_i}{2}$

Example:

Definition (Graph transitivity)

The **transitivity** T of a graph $G = (X, E)$ is the proportion of **transitive** triads, i.e., triads which are subgraphs of **triangles**. This proportion can be computed as follows:

$$T = 3\frac{n_\Delta}{n_\wedge},$$

where n_Δ is the number of triangles in G, and $n_\wedge$ is the number of triads.