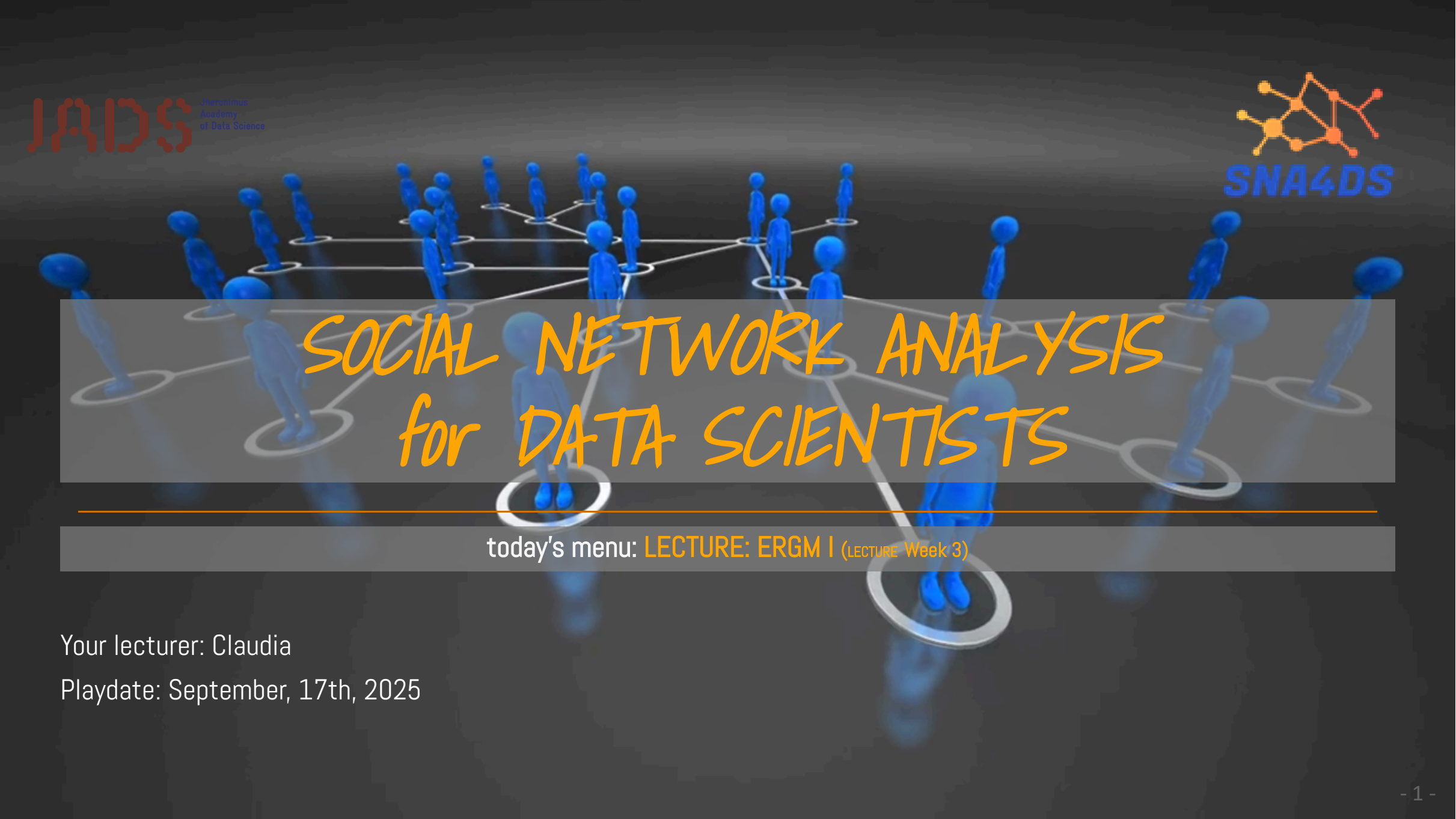


SOCIAL NETWORK ANALYSIS for DATA SCIENTISTS

today's menu: LECTURE: ERGM I (LECTURE Week 3)

Your lecturer: Claudia

Playdate: September, 17th, 2025

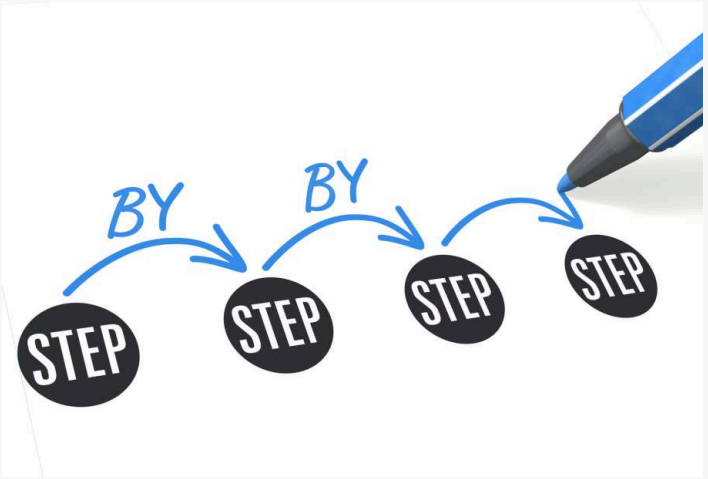
Goal for today:

Introducing Exponential Random Graph models

These models are going to stay with us for a while

yes, they are in the exam

yes, you MUST use them in the project



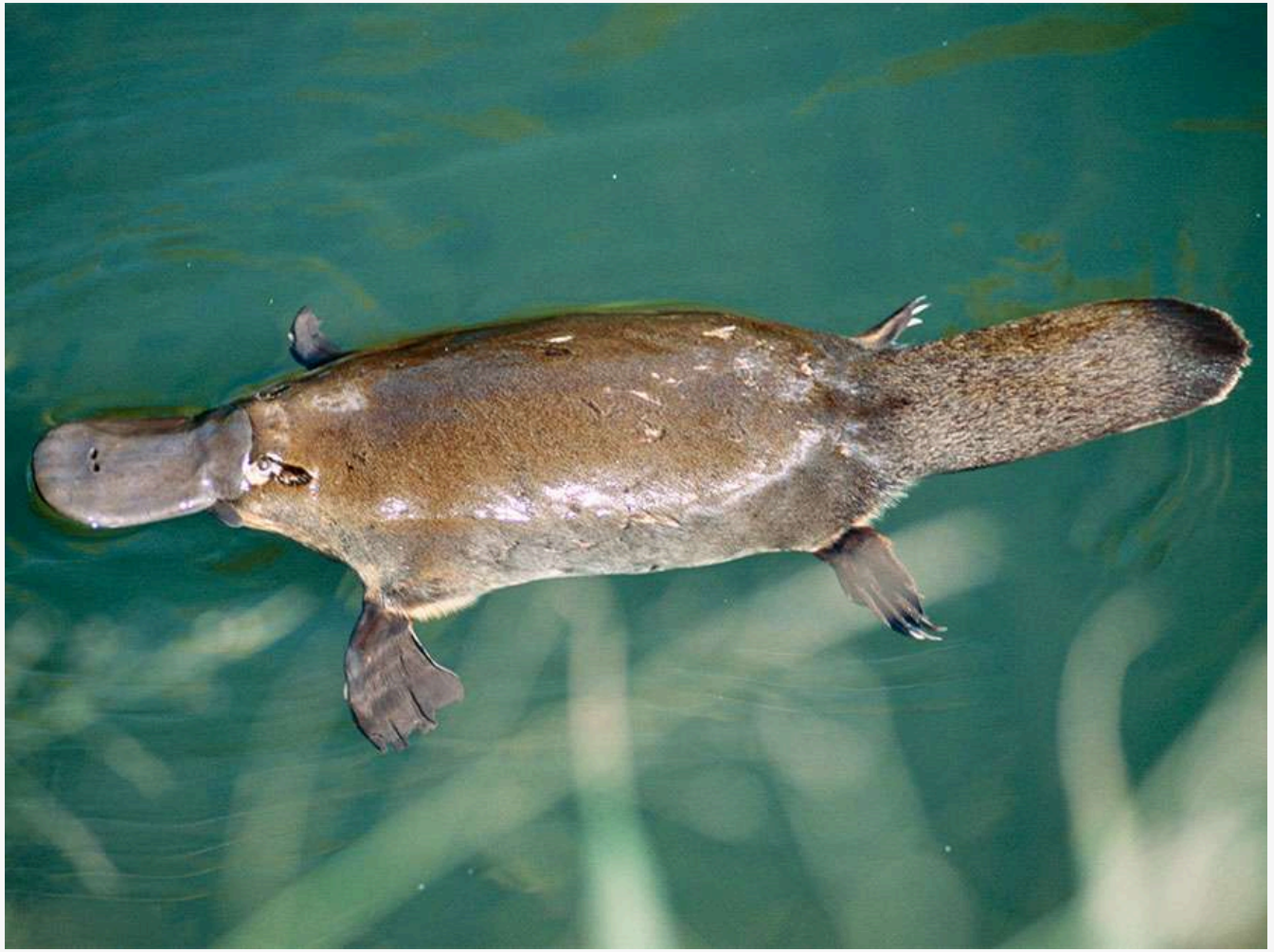
Today we cover

- what the model is (for)
- theory

So that we make sure that you create a solid connection :)

In the following weeks we will touch data, R package, code, so on and so forth

Disclaimer!



It seems weird, but when you get used to it, it is just another animal :)

Menu' for today

SNA4DS

1. Causal Inference Mindset
2. A matter of outcome variable
3. Evolving toward the ERGM
4. A matter of predictors
5. How the model works
6. More about the predictors
7. Let's observe a case!
8. Software info

1. Causal Inference Mindset

NETWORKS

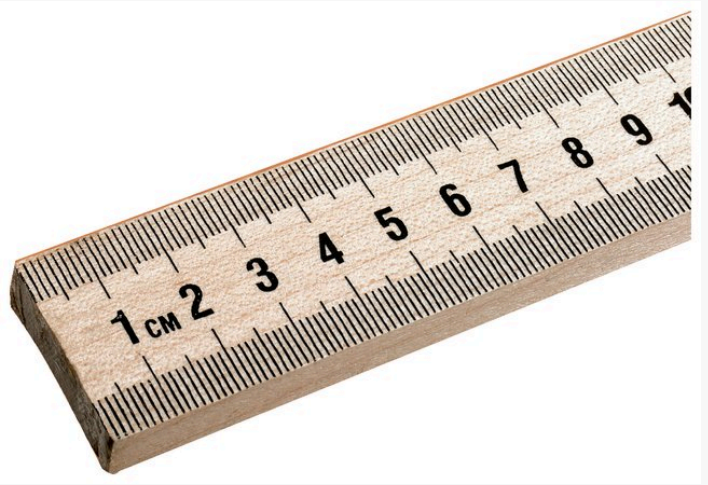
We

observed them

conceptualized them

described them

analyzed their parts and components



STILL

Are we truly understanding them?

Is a description always enough?

WE CAN DO MORE

Understanding Networks

SNA4DS

Why is a network the way it is?

- Networks conceptualize relationships of many different kinds
- Relationships are not random

  you don't have random friends

  you don't marry someone at random

  two atoms are not connected at random

  streets don't connect random parts of the city

Describing these relationships is not enough.

What causes this specific connections?

Causal Inference for networks

Do you remember causal inference?

You had a class with Claudia in the Bootcamp!

Prediction VS Inference

Inference = Deducing

You already know that you can deduce the causes that generated a network

You had a class with Roger in SNA4DS --- statistical models for networks!

Inductive VS Deductive

Causal inference RECAP

- for IID data (non network)

  OLS

  T-test

  GLM -- Logistic regressions

- for networks

  NetLogit

  MRQAP

  LNAM

Today we add ERGMs to the list

ERGMs are a class of network models for causal inference

They are a class of models to deduce the cause of an observed relational phenomenon that we observe/conceptualize as a network.

Causal inference RECAP 2a

Correlation and causation RECAP

2 variables: body weight, gender

- correlation: body weight & gender (order does not matter)
- causation

formula for option 1: body weight ~ gender
formula for option 2: gender ~ body weight

Which one makes sense?

Causal inference RECAP 2b

If 0 is male and 1 is female

gender ~ body weight

our hypothesis is:

- If a pregnant woman eats a lot is going to give birth to a girl
  You do realize that this makes no sense, right?

body weight ~ gender

Our model can only test the other hypothesis:

- Men tend to be bigger than women
  OBVIOUSLY the only Hypothesis you can test here

Causality Matters!

2. A matter of outcome variable

Explaining network structure

ERGMS, as any other statistical model, work with the hypothesis testing (causal) mindset and have:

Outcome variable <- what do we want to explain

Predictors <- what we believe could explaining it

The hypothesis is a sentence that explains how you believe that predictor affects the outcome

We are explaining network structure, right?

Our outcome variable MUST be a network

How do you put a network as an outcome variable?

A network describes two things

relationship --- edge [1]

no relationship --- no edge [0]

Does this sound familiar to you?

How do you put a network as an outcome variable?

SNA4DS

A network describes two things

relationship --- edge [1]

no relationship --- no edge [0]

Does this sound familiar to you?

A logistic regression

Statistical Mindset RECAP 2.0

- In non-network regression models (OLS, GLM, ...) we test whether one or more covariates/predictors predict the outcome variable (coef, p-value).
- This is intended to test whether an hypothesized causal relationships exists
- Did the predictor(s) affect the outcome? (first: theory; second: math)
- Our p-value is the probability that we can exclude that a certain effect is random
- Our p-value is NOT a probability of the hypothesis being true, it is the conditional probability of being able to reject the null hypothesis given the data
    - We repeat the experiment 1000 times.
    - We compare our real data to the 1000 experimental results
    - If real data = 1000 repetitions it happened by chance
    - If real data != 1000 repetitions we can exclude that what we observe happened by chance in our data.

Statistical Mindset - Moving on

It seems the same thing that we are doing here, right?

Can we use a Logistic regression for networks?

Statistical Mindset - Moving on

It seems the same thing that we are doing here, right?

Can we use a Logistic regression for networks?

NO

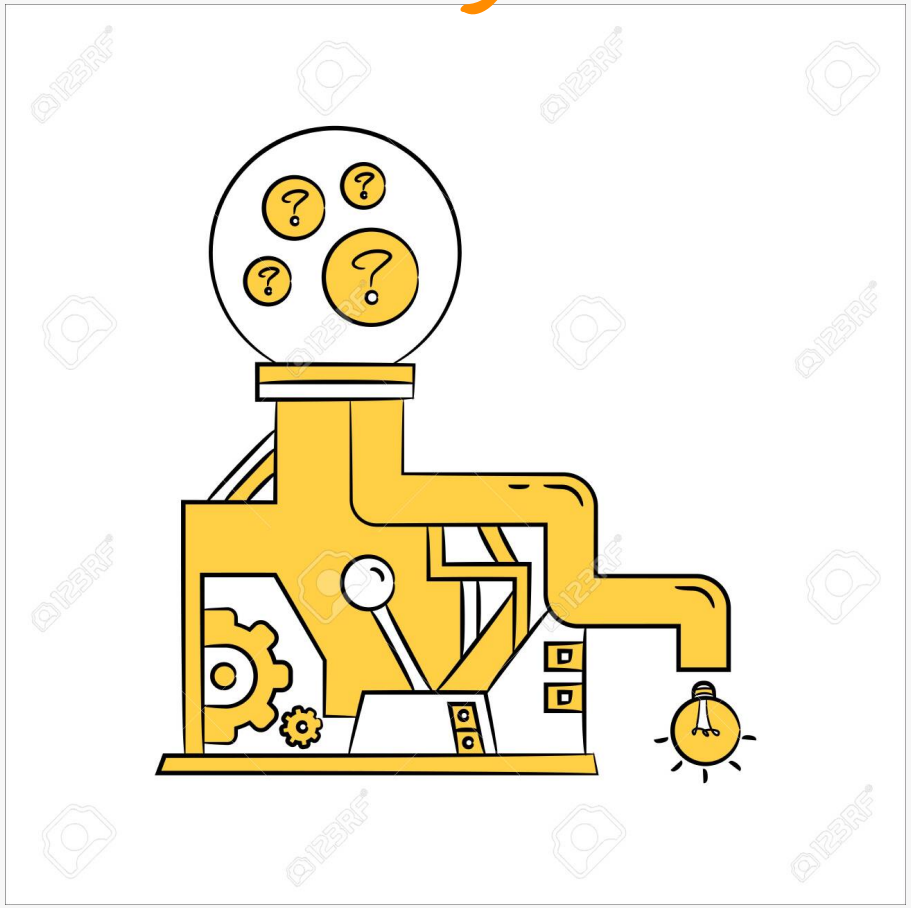
Assumption Violation

Observations must be Independent and Identically Distributed (IID for friends)

If you use a non network model your errors estimations will be incorrect!

...explaining a network takes more work than explaining a regular variable!

We need a new solving problem machine!



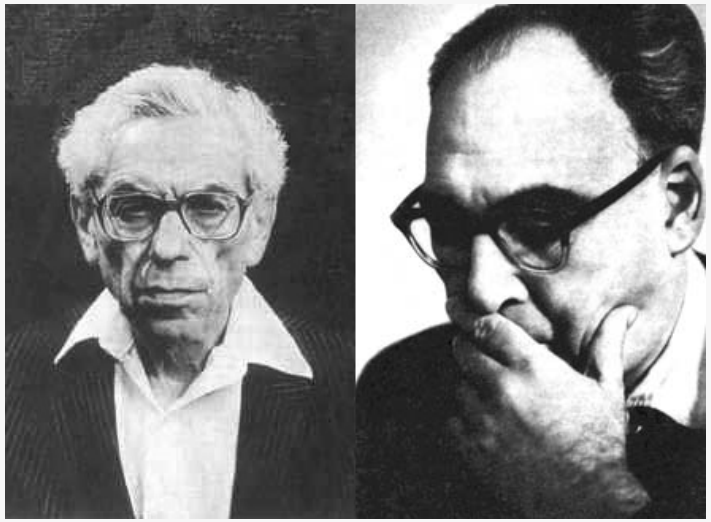
The ERGM

3. Evolving toward the ERGM

Erdos Renyi Game

they introduced the "game" of creating random network given fixed parameters

- N - number of nodes
- E - number of edges (or probability of edge formation)



A very simple ERGM

Looking at the structure of the network

What is the probability that we observe a network with that specific connections out of all the possible combinations?

In probability terms this is:

$$logP_{ij}(0/1) = \lambda_{ij}$$

Where

- $P_{ij}(0,1)$ is the probability of having or not having an edge between each pair of nodes.
- λ_{ij} a parameter that expresses a rate for edge propagation (in our case N of edges)

We compare an observed network against random networks generated with the same parameters.

Research design

General example

RQ: Given the number of nodes in my network, how likely it is that my number of edges is random?

H1: The number of edges I observe is not random

H0: The number of edges is random

Specific example

RQ: Given that in this class there are 80 students, how likely it is that they are friends with anyone?

H1: I observe THIS number of friendships between pairs of students given by the fact that they get along

H0: I observe THIS number of friendships between pairs of students and there is no reason for it

More complex example

Consider all the students at JADS as nodes

Let's say that JADS organizes an Hackathon with 5K euros at stake

You have to form groups to work together every day

If you study together there is an edge between each pair of you

I map down the study-together dynamics two days before the Hackathon

If I compare that network to random ones, do you think my observed one will show some particular features?

More complex example

Consider all the students at JADS as nodes

Let's say that JADS organizes an Hackathon with 5K euros at stake

You have to form groups to work together every day

If you study together there is an edge between each pair of you

I map down the study-together dynamics two days before the Hackathon

If I compare that network to random ones, do you think my observed one will show some particular features?

Yes. Connections will be different than random, they are driven by a purpose (e.g., everyone wants to connect with good coders etc...)

Let's use more predictors (P1 model)

Let's do something a little more juicy, shall we?

We have a directed network such as Instagram followers.

You might follow me, but i might not follow you.

We can observe:

- senders (followers) A --> B
- receivers (followed) A <-- B
- mutual (followers & receivers together) A <--> B

P(0, 0) = no edge

P(1, 0) = send an edge

P(0, 1) = receive an edge

P(1, 1) = mutual

This model in math [P1 model]

$$logP_{ij}(1/0) = \lambda_{ij} + \alpha_i + \beta_j + \alpha_j + \beta_i + 2\theta + \rho_{ij}$$

λ_{ij} probability of the density (N of edges) ER

α_i sender i

β_j receiver j

α_j sender j

β_i receiver i

$2\theta + \rho_{ij}$ mutual

We want to observe these parameters and see whether they are or not random in the network

Expanding our horizons

If we can measure density, sender, receiver, and mutual,

Why not the rest?

You know how to measure networks

Dyad Census

Triad Census

...

Are these effects random or are these effect driven by some reason?

In the ERGM we can estimate the probability of particular descriptive measures that we extract from the network to be non-random (reject the null)

-- we still don't know the cause, but we can assume it is a good one.

4. A matter of predictors

Two main Different types of predictors

Exogenous: Outside

Endogenous: Inside

Endogenous

Effects occurring within the outcome network of interest

(what happens inside the house in the picture -- inside the network)

We already saw

- edge
- sender
- receiver
- mutual
- ...
- other measures of interest

Exogenous

Effects occurring outside of the outcome network

(like the hurricane)

Intuitively, this type of variables correspond to the extra information that is not naturally embedded in the network - They are outside of the network

For example:

If we want to explain the reason why JADS students might play football together using the variable "How much do you like football" 1-5 might already help, right?

5. How the model works

Exponential Random Graph Model (P*)

$$P(N,\theta) = \frac{exp\{\theta^T h(N)\}}{\sum_{N^* \in \mathcal{N}} exp\{\theta^T h(N^*)\}^T}$$

Exponential Random Graph Model (P*)

$$P(N,\theta) = \frac{exp\{\theta^T h(N)\}}{\sum_{N^* \in \mathcal{N}} exp\{\theta^T h(N^*)\}^T}$$

- $P(N,\theta)$ - Probability of observing this real Network, with these coefficients

Exponential Random Graph Model (P*)

$$P(N,\theta) = \frac{exp\{\theta^T h(N)\}}{\sum_{N^* \in \mathcal{N}} exp\{\theta^T h(N^*)\}^T}$$

- $P(N,\theta)$ - Probability of observing this real Network, with these coefficients
- $h(N)$ - Using a combination of network statistics taken from the observed network

Exponential Random Graph Model (P*)

$$P(N,\theta) = \frac{exp\{\theta^T h(N)\}}{\sum_{N^* \in \mathcal{N}} exp\{\theta^T h(N^*)\}^T}$$

- $P(N,\theta)$ - Probability of observing this real Network, with these coefficients
- $h(N)$ - Using a combination of network statistics taken from the observed network
- θ^T - Estimating this/these parameter(s)

Exponential Random Graph Model (P*)

$$P(N,\theta) = \frac{exp\{\theta^T h(N)\}}{\sum_{N^* \in \mathcal{N}} exp\{\theta^T h(N^*)\}^T}$$

- $P(N,\theta)$ - Probability of observing this real Network, with these coefficients
- $h(N)$ - Using a combination of network statistics taken from the observed network
- θ^T - Estimating this/these parameter(s)
- N^* - Confronting the real network with all the permutations that are possible

Exponential Random Graph Model (P*)

$$P(N,\theta) = \frac{exp\{\theta^T h(N)\}}{\sum_{N^* \in \mathcal{N}} exp\{\theta^T h(N^*)\}^T}$$

- $P(N,\theta)$ - Probability of observing this real Network, with these coefficients
- $h(N)$ - Using a combination of network statistics taken from the observed network
- θ^T - Estimating this/these parameter(s)
- N^* - Confronting the real network with all the permutations that are possible
- $\sum_{N^* \in \mathcal{N}} exp\{\theta^T h(N^*)\}^T$ normalizing coefficient: Exponential function of the summation of all possible permutation weighted for all the coefficients estimated with the given data

Exponential Random Graph Model (P*)

$$P(N,\theta) = \frac{exp\{\theta^T h(N)\}}{\sum_{N^* \in \mathcal{N}} exp\{\theta^T h(N^*)\}^T}$$

- $P(N,\theta)$ - Probability of observing this real Network, with these coefficients
- $h(N)$ - Using a combination of network statistics taken from the observed network
- θ^T - Estimating this/these parameter(s)
- N^* - Confronting the real network with all the permutations that are possible
- $\sum_{N^* \in \mathcal{N}} exp\{\theta^T h(N^*)\}^T$ normalizing coefficient: Exponential function of the summation of all possible permutation weighted for all the coefficients estimated with the given data

*Permutation: check all the possible combination that exist between each pair of nodes

6. More about the predictors

Model Terms or Effects

SNA4DS

IID linear model (GLM): We can test hypotheses with predictors and controls

You get an estimate and check whether the outcome variable

increases or decreases

When predictor/control variates

These predictors/controls are always information about something

You already saw that, but with ERGMs we can be more flexible.

We can customize model terms as much as we like! including both information (exogenous) and network structure (endogenous)

Plug in effects in the equation

Many things that you can observe in a network can be inserted in an ERGM

$h(N)$ - Using a combination of network statistics

We can plug in effects in the equation here (see the full equation above).

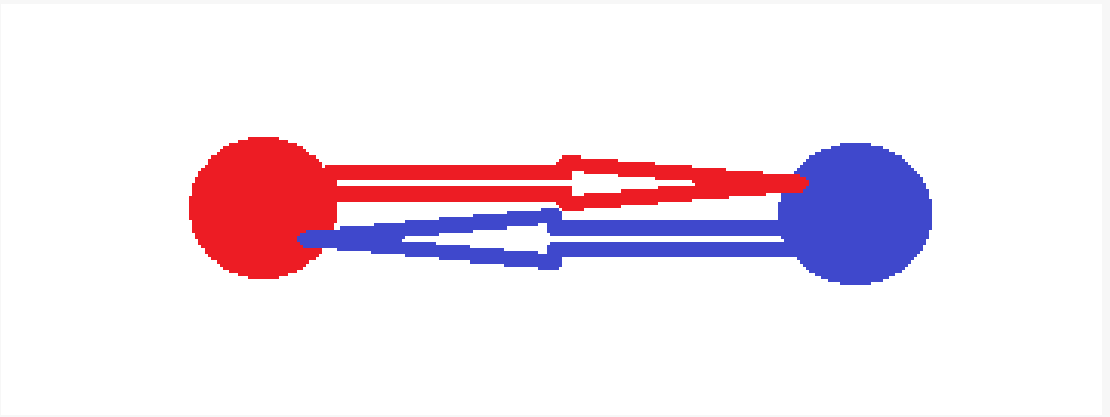
Example: Mutual

In the P1 model we already saw the mutual term

Reciprocity: If I follow you on Instagram, what is the probability that you will follow me back?

In Math

$h = \sum_{i<j} N_{ij} N_{ji}$



ERGM with Reciprocity effect

SNA4DS

substitute the reciprocity formula in the generic parameter $h(N)$

$$P(N,\theta) = \frac{exp\{\theta^T(\sum_{i<j} N_{ij}N_{ji})\}}{\sum_{N^* \in \mathcal{N}} exp\{\theta^T(\sum_{i<j} N^*_{ij}N^*_{ji})\}^T}$$

If you insert more than one effect you keep adding them up - here it is edges + mutual

$$P(N,\theta) = \frac{\exp\{\theta^T(\sum_{i\neq j} N_{ij} + \sum_{i<j} N_{ij}N_{ji})\}}{\sum_{N^* \in \mathcal{N}} \exp\{\theta^T(\sum_{i\neq j} N^*_{ij} + \sum_{i<j} N^*_{ij}N^*_{ji})\}^T}$$

What you ultimately want is to reconstruct a list of parameters that explain in the best possible way what is going on with your data.

Same as tailors with their client!

Sim city

In other words, it is like playing Sim City. It's just that you don't see the graphic interface, but model parameters :)

Recap

Outcome variable: Always a Network

Predictors: Exogenous and Endogenous

Effects: for both kind of predictors, several effects to look at

Prediction: Linear Model

Specification: Theory driven (Research question, hypothesis testing)

How it works:

- We specify a model that we believe generates the synthetic reality we care for
- The model expresses a probability distribution that generate networks that behave as the observed one
- We compare real data to the generated data
- We can make some conclusions on what caused the relationships observed in the network (test the hypothesis)

How does this run? We will talk about the algorithm to estimate these models in the following lessons.

7. Let's observe a case!

Data

We collected network data in a school since we are interested in understanding the dynamics of friendship.
We have three nodal attributes

- their gender
- their age
- who is a smoker

Questions and Hypothesis

(thinking in regressions mindset -- causal mindset)

RQ: What are the drivers of friendship in this school?
After doing some research we believe that young people are

- more likely to make friends with people of the same gender
- more likely to make friends with people closer to their age
- smokers are more likely to be friends with other smokers

Questions and Hypothesis

(thinking in regressions mindset -- causal mindset)

RQ: What are the drivers of friendship in this school?
After doing some research we believe that young people are

- more likely to make friends with people of the same gender
- more likely to make friends with people closer to their age
- smokers are more likely to be friends with other smokers

These hypotheses all refer to exogenous predictors!

Exogenous effects

- more likely to make friends with people of the same gender

HOMOPHILY

Exogenous effects

- more likely to make friends with people of the same gender

HOMOPHILY

- more likely to make friends with people closer to their age

ABSOLUTE DIFFERENCE IN AGE

Exogenous effects

- more likely to make friends with people of the same gender

HOMOPHILY

- more likely to make friends with people closer to their age

ABSOLUTE DIFFERENCE IN AGE

- smokers are more likely to be friends with other smokers

HOMOPHILY

Questions and Hypothesis (thinking ergm)

We know that friendship in a school does not have homogeneous patterns (== to random)

- There are some people that are more popular than others
- There are some people that are more social than others

Questions and Hypothesis (thinking ergm)

We know that friendship in a school does not have homogeneous patterns (== to random)

- There are some people that are more popular than others
- There are some people that are more social than others

These hypotheses all refer to endogenous predictors!

Endogenous effects

- There are some people that are more popular than others

POPULARITY -- preferential attachment (Barabasi and Albert)

Receiving ties

Endogenous effects

- There are some people that are more popular than others

POPULARITY -- preferential attachment (Barabasi and Albert)

Receiving ties

- There are some people that are more social than others

SOCIALITY

Sending ties

Model specification

After we did this research design we would know what to insert in our model

In PROTO-CODE, the model would be something like that

school ~ HOMOPHILY (Gender) + ABSOLUTE DIFFERENCE (Age) + HOMOPHILY (Smokers) + POPULARITY + SOCIALITY

For this round I leave the math to you since this starts to be quite long :)

8. Software info

statnet

Suite of R packages for the management, exploration, statistical analysis, simulation and visualization of network data.

http://statnet.org/

It takes objects of class `networks`

It does NOT take objects of class `igraph`

And now it is time for tea!