A short course on

Social Network Analysis

Lecture 1

Antonio Calcagnì

⟨antonio.calcagni@unipd.it⟩

Master in Data Science per il Welfare

University of Salento

AY 2023/24

Outline

1 Modeling network data

2 Mathematical Network models

3 Statistical network models

Outline

1. Modeling network data

2. Mathematical Network models

3. Statistical network models

Modeling network data

Descriptive statistics offer a static overview of network properties like centrality measures and clustering coefficients but lack predictive power and the ability to test hypotheses about network dynamics and structure.

They also do not incorporate node attributes or control for confounding factors, limiting their utility for comprehensive network analysis.

Modeling network data

In contrast, **network models** uncover underlying network processes, predict future states, and allow for formal hypothesis testing. These models integrate node attributes, simulate scenario changes, and reveal hidden structures such as community patterns.

They provide actionable insights across fields like epidemiology and sociology, making them essential for understanding and leveraging complex network dynamics.

Modeling network data

Network models have a **rich historical development**. Sociologists and statisticians made significant strides during the 1970s and 1980s, leading to the creation of substantial databases and the introduction of exponential random graph models and related techniques by the early 1990s.

Physicists and computer scientists entered the field later but expanded the range of models and methodologies. They focused on larger networks and more intricate forms of data analysis, enhancing the field's depth and breadth.

Modeling network data

A **model** for a network graph $\mathcal{G}$ is a collection of graphs $\mathcal{H} = \{\mathcal{G}_1, \ldots, \mathcal{G}_g, \ldots, \mathcal{G}_G\}$ equipped with an indexed probability distribution $\mathbb{P}_{\boldsymbol{\theta}}(\mathcal{G})$ over $\mathcal{H}$, i.e.:

$$\{\mathbb{P}_{\boldsymbol{\theta}}(\mathcal{G}), \mathcal{G} \in \mathcal{H}, \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$$

The flexibility of the model is given by the choice of $\mathbb{P}_{\boldsymbol{\theta}}(\mathcal{G})$, with simplest models letting $\mathbb{P}_{\boldsymbol{\theta}}(\mathcal{G}) = \mathcal{U}(\mathcal{G}; \alpha, \beta)$ and more complex models assuming $\mathbb{P}_{\boldsymbol{\theta}}(\mathcal{G})$ be a member of the Exponential Family of distributions.

To further characterize the network model, usually *constraints* are used to define the ensemble $\mathcal{H}$, e.g. by fixing some graph properties like N_v or the degree distribution $\{f_d\}_{d>0}$.

Modeling network data

Didactically speaking, network models can be broadly classified into two classes:

Mathematical models

Statistical models

Modeling network data

Didactically speaking, network models can be broadly classified into two classes:

Mathematical models: Define graph models from a theoretical and top-down perspective, with a special emphasis on the class of reasonable models for many empirical phenomena. Examples: Erdos-Renyi, Scale-free, Small-world.

Statistical models

Modeling network data

Didactically speaking, network models can be broadly classified into two classes:

Mathematical models: Define graph models from a theoretical and top-down perspective, with a special emphasis on the class of reasonable models for many empirical phenomena. Examples: Erdos-Renyi, Scale-free, Small-world.

Statistical models: More closed to the statistical approach of modeling phenomena, define graph models that are estimable from the data and tested by also including exogenous variables or covariates. Examples: ERG, gERG, LS models.

Outline

1 Modeling network data

2 Mathematical Network models

3 Statistical network models

Mathematical network models

Random graph models

The standard **Erdos-Renyi model** specifies a collection $\mathcal{H}$ of graphs $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ all of them with the same number of nodes N_v and edges N_e (constraint). In addition, it assigns to each graph a *uniform* probability:

$$\mathbb{P}_{\boldsymbol{\theta}}\left(\mathcal{G}\right) = \binom{N_v/2}{N_e}^{-1}$$

In the **Gilbert's formulation** $\mathcal{H}$ can be obtained by assigning e independently to each distinct pair $\{u, v\}$ with probability $p \in (0, 1)$. It assigns to each graph a *Bernoulli* probability:

$$\mathbb{P}_{\boldsymbol{\theta}}\left(\mathcal{G}\right) = p^{N_e}(1-p)^{\binom{N_v}{2}-N_e}$$

Mathematical network models

Random graph models

Instead of constraint the ensemble $\mathcal{H}$ by fixing $N_v^{(g)}$ and $N_e^{(g)}$ for $g \in \{1, \dots, G\}$, one can constraint the degree sequence $\{d_{(1)}, \dots, d_{(N_v)}\}$ to be the same over the graphs. This leads to a **generalized random graph model**.

There are also variants where the degree distribution $\{f_d\}_{d>0}$ is kept fixed over the graphs instead of the degree sequence (e.g., Molly and Reed's model).

Mathematical network models

Random graph models

Instead of constraint the ensemble $\mathcal{H}$ by fixing $N_v^{(g)}$ and $N_e^{(g)}$ for $g \in \{1, \dots, G\}$, one can constraint the degree sequence $\{d_{(1)}, \dots, d_{(N_v)}\}$ to be the same over the graphs. This leads to a **generalized random graph model**.

There are also variants where the degree distribution $\{f_d\}_{d>0}$ is kept fixed over the graphs instead of the degree sequence (e.g., Molly and Reed's model).

Note: RGMs often serve as a way to define the *null distribution* in statistical hypothesis tests in network data analysis.

Mathematical network models

Graph models based on mechanisms

Beyond the family of RGMs, there exist network graph models which are built so as to mimic certain real-world mechanisms, for instance by reproducing the network clustering properties or certain growth dynamics.

Mathematical network models

Graph models based on mechanisms

The **scale-free model** characterizes $\{f_d\}$ of a graph $\mathcal{G}$ using the Power-Law distribution $\mathbf{d}_{\mathcal{G}} \sim d^{-\lambda}$.

The intuition here is that a few nodes show a very high degree (*hubs*) and many nodes show relatively few connections instead. These networks are dominated by a small number of highly connected nodes (hubs) that play a crucial role in information propagation and network resilience.

Mathematical network models

Graph models based on mechanisms

Another well-known model is the **preferential-attachment**, which mimics the *rich get richer* phenomenon. In this process, new nodes in a network preferentially connect to existing nodes that already have a high degree. Nodes with higher degrees attract more new connections, reinforcing their centrality in the network.

Preferential attachment results in the emergence of *hubs*, i.e. nodes with a disproportionately high number of connections. These networks typically exhibit scale-free characteristics.

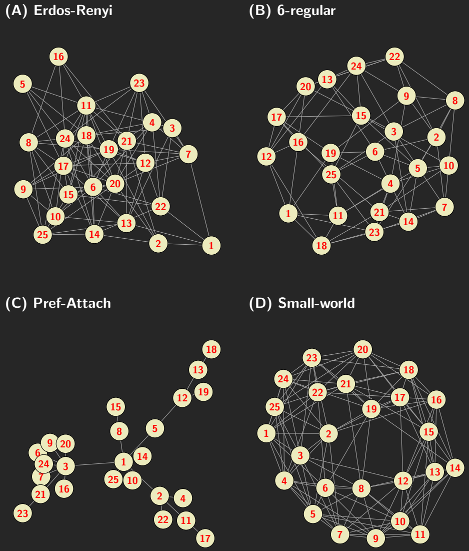
Mathematical network models

Graph models based on mechanisms

To formalize the *six-degree separation* phenomenon (i.e., most of the nodes are six or fewer connections away from each other), the **small-world model** can be used. The network is built from a regular grid where nodes are iteratively clustered by progressively reducing the path lengths (*rewiring*).

Small-world networks model many real-world systems, facilitating studies on information diffusion, disease spread, and social interactions.

Mathematical network models



(A) Erdos-Renyi

(B) 6-regular

(C) Pref-Attach

(D) Small-world

Mathematical network models

Assess significance in network graphs

Although too simplistic for real statistical modeling, the graph models described so far play a useful role in statistical hypothesis testing.

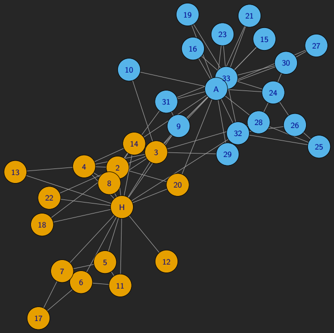
Suppose we have derived a graph $\mathcal{G}^{\text{obs}}$ from a set of observations and we are interested in one or more of its characteristics $\eta(\mathcal{G})^{\text{obs}}$ like the number of triads or the centrality. To establish if they are somehow unusual or unexpected (i.e., *significant*), we can simulate the null distribution of the characteristics being studied and quantify their unlikeness:

$$\mathbb{P}_{\eta(\mathcal{G})}(t) = \frac{|\{\mathcal{G} \in \mathcal{H} : \eta(\mathcal{G}) \leq t\}|}{|\mathcal{H}|}$$

As usual, comparing $\mathcal{G}^{\text{obs}}$ with this distribution allows for quantifying the evidence against the null hypothesis.

Mathematical network models

Assess significance in network graphs



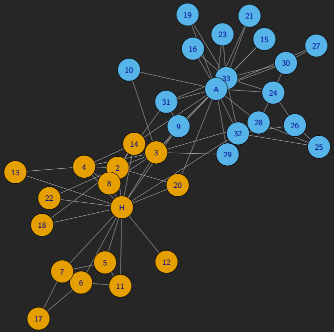
Zachary's karate club

In the Zachary's karate network we have found a *moderate* level of clustering $C_T = 0.25$.

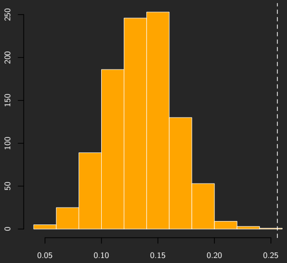
With no other comparable information (e.g., networks of similar clubs), it is difficult to conclude **whether or not** a value of 0.25 is in any sense unexpected **given this type** of network graphs.

Mathematical network models

Assess significance in network graphs



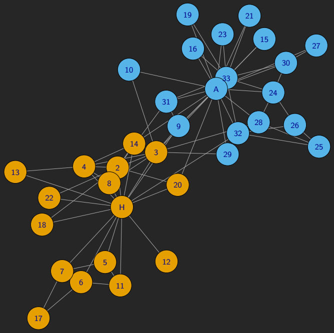
Zachary's karate club



Null distribution of C_T under the hypothesis that $\mathcal{G}^{\text{obs}}$ has been drawn from a Erdos-Renyi model with $N_v = 34$ and $N_e = 78$ (no. of replicates $B = 1000$). Note that the dotted white line indicate the observed $C_T = 0.25$.

Mathematical network models

Assess significance in network graphs



Zachary's karate club

In the Zachary's karate network we have found a *moderate* level of clustering $C_T = 0.25$.

We conclude that the Zachary's network shows markedly greater transitivity than other random graphs of comparable magnitude or connectivity.

Outline

1 Modeling network data

2 Mathematical Network models

3 Statistical network models

Statistical network models

Depending on the research question one is tacking, there are several statistical models that can be adopted to study network graphs:

- Exponential Random Graph (ERG): a kind of Generalized Linear Model for network data
- Network Block (NB): similarly to mixture models, they allow for directly modeling subgroups or blocks
- Latent Network (LN): allow for including latent variables in the graph formation process
- Temporal Network (TN): allow for modeling structures evolving over time

Statistical network models

Depending on the research question one is tacking, there are several statistical models that can be adopted to study network graphs:

- Exponential Random Graph (ERG): a kind of Generalized Linear Model for network data
- Network Block (NB): similarly to mixture models, they allow for directly modeling subgroups or blocks
- Latent Network (LN): allow for including latent variables in the graph formation process
- Temporal Network (TN): allow for modeling structures evolving over time

→ We will be focusing on ERG models only

Note: It is a simple and basic introduction to the topic.

Statistical network models

ERG model

Consider the lower triangular part of the adjacency matrix $\mathbf{y} = \text{vec}(\mathbf{Y}_\triangle)$ for an undirected graph $\mathcal{G}$. This is a vector of $N_v(N_v - 1)/2$ elements, with $y_{ij} \in \{0, 1\}$ indicating whether an edge exists between the vertices i and j.

The ERG models is of the form:

$$\mathbb{P}(\mathbf{y} = y; \boldsymbol{\theta}) \propto \exp\left(\boldsymbol{\theta}_1^T \eta(\mathbf{y}) + \boldsymbol{\theta}_2^T h(\mathbf{z}, \mathbf{y})\right)$$

where:

- $\eta(\mathbf{y}) \in \mathbb{R}^p$ is a $p \times 1$ vector of *network statistics* (endogenous information)
- $h(\mathbf{z}, \mathbf{y}) \in \mathbb{R}^p$ is a $p \times 1$ vector of *network attributes* (exogenous information) depending on both $\mathbf{y}$ and $\mathbf{z}$
- $\boldsymbol{\theta}_1 \cup \boldsymbol{\theta}_2 = \boldsymbol{\theta} \in \mathbb{R}^{2p}$ is the vector of model parameters

Statistical network models

ERG model

$$\mathbb{P}(\mathbf{y} = y; \boldsymbol{\theta}) \propto \exp\left(\boldsymbol{\theta}_1^T \boxed{\eta(\mathbf{y})} + \boldsymbol{\theta}_2^T h(\mathbf{z}, \mathbf{y})\right)$$

The quantity $\eta(\mathbf{y})$ is built so as to summarize the (endogenous) complexity of network structure, such as the number of edges, the number of k-stars, the number of triangles.

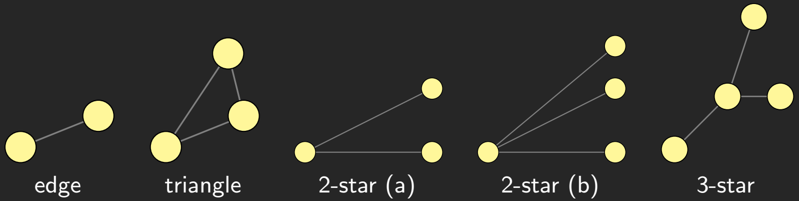
Based on the Markovian assumption, the aim is to capture a form of **local dependency**: two edges $\{i, j\}$ and $\{i', j'\}$ with $j \neq j'$ are dependent whenever they share a vertex v conditioned on the status of the other possible edges of the graph (**Markov graph**).

Statistical network models

ERG model

$$\mathbb{P}(\mathbf{y} = y; \boldsymbol{\theta}) \propto \exp\left(\boldsymbol{\theta}_1^T \boxed{\eta(\mathbf{y})} + \boldsymbol{\theta}_2^T h(\mathbf{z}, \mathbf{y})\right)$$

Common types of endogenous network characteristics:

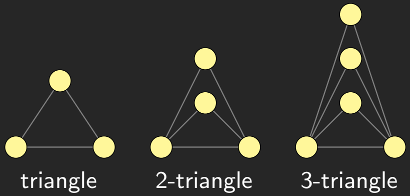


Statistical network models

ERG model

$$\mathbb{P}(\mathbf{y} = y; \boldsymbol{\theta}) \propto \exp\left(\boldsymbol{\theta}_1^T \boxed{\eta(\mathbf{y})} + \boldsymbol{\theta}_2^T h(\mathbf{z}, \mathbf{y})\right)$$

Common types of endogenous network characteristics:



Statistical network models

ERG model

$$\mathbb{P}(\mathbf{y} = y; \boldsymbol{\theta}) \propto \exp\left(\boldsymbol{\theta}_1^T \boxed{\eta(\mathbf{y})} + \boldsymbol{\theta}_2^T h(\mathbf{z}, \mathbf{y})\right)$$

There are several combinations of network characteristics that can b used to capture local dependencies. Some of them have a direct interpretation such as:

- Reciprocity (edge)
- Transitivity, i.e. the friends of my best friends are also my friends (triangles)
- Popularity (stars)
- Close-knit groups (k-triangles)

Statistical network models

ERG model

$$\mathbb{P}(\mathbf{y} = y; \boldsymbol{\theta}) \propto \exp\left(\boldsymbol{\theta}_1^T \boxed{\eta(\mathbf{y})} + \boldsymbol{\theta}_2^T h(\mathbf{z}, \mathbf{y})\right)$$

These structure can be counted by using weighted summary statistics:

$$\text{AKS}_\lambda(\mathbf{y}) = \sum_{k=2}^{N_v - 1} (-1)^k \frac{\tau_{\star(k)}}{\lambda^{k-2}} \qquad \textit{alternating k-star}$$

where

- $\tau_{\star(k)}$ is the number of k-star forms in the observed network
- λ scales the statistic (a higher value would increase the likelihood of promoting $\star$-like structures)

Statistical network models

ERG model

$$\mathbb{P}(\mathbf{y} = y; \boldsymbol{\theta}) \propto \exp\left(\boldsymbol{\theta}_1^T \boxed{\eta(\mathbf{y})} + \boldsymbol{\theta}_2^T h(\mathbf{z}, \mathbf{y})\right)$$

These structure can be counted by using weighted summary statistics:

$$\text{AKT}_\lambda(\mathbf{y}) = 3\tau_\triangle(1) + \sum_{k=2}^{N_v-2} (-1)^{k+1} \frac{\tau_{\triangle(k)}}{\lambda^{k-1}} \qquad \textit{alternating } k\textit{-triangles}$$

where

- $\tau_{\triangle(k)}$ is the number of k-triangles forms in the observed network
- λ scales the statistic (a higher value would increase the likelihood of promoting $\triangle$-like structures)

Statistical network models

ERG model

$$\mathbb{P}(\mathbf{y} = y; \boldsymbol{\theta}) \propto \exp\left(\boldsymbol{\theta}_1^T \eta(\mathbf{y}) + \boldsymbol{\theta}_2^T \boxed{h(\mathbf{z}, \mathbf{y})}\right)$$

The quantity $h(\mathbf{z}, \mathbf{y})$ is built so as to summarize the exogenous effects of **node**-level or **edge**-level covariates on the graph structure. The latter are also known as dyadic effects.

We can include: continuous and categorical variables (*main effects*), homophily or similarity (*second-order effects*), edge-matched covariates (*dyadic effects*), and many others.

Statistical network models

ERG model

$$\mathbb{P}(\mathbf{y} = y; \boldsymbol{\theta}) \propto \exp(\theta N_e)$$

The model with network statistic being defined only in terms of number of edges often serves as a **baseline model** as it loses the Markov property (i.e., it is a simple Bernoulli or Erdos-Renyi model). It can be used as starting point for testing more complex models.

Statistical network models

ERG model

$$\mathbb{P}(\mathbf{y} = y; \boldsymbol{\theta}) \propto \exp\left(\boldsymbol{\theta}_1^T \eta(\mathbf{y}) + \boldsymbol{\theta}_2^T h(\mathbf{z}, \mathbf{y})\right)$$

The model parameters $\boldsymbol{\theta}$ are usually based on MCMC-based maximum likelihood approximation. However, the inferential problem based on such approximation does not guarantee asymptotic results for confidence intervals.

Statistical network models

ERG model

$$\mathbb{P}(\mathbf{y} = y; \boldsymbol{\theta}) \propto \exp\left(\boldsymbol{\theta}_1^T \eta(\mathbf{y}) + \boldsymbol{\theta}_2^T h(\mathbf{z}, \mathbf{y})\right)$$

Finally, the **goodness-of-fit** of the ERG model can be assessed using simulation-based diagnostics and discrepancy measures. We will see more about that during the practical session.

Statistical network models

ERG model

$$\mathbb{P}(\mathbf{y} = y; \boldsymbol{\theta}) \propto \exp\left(\boldsymbol{\theta}_1^T \eta(\mathbf{y}) + \boldsymbol{\theta}_2^T h(\mathbf{z}, \mathbf{y})\right)$$

The **parameter interpretation** is similar to the GLMs case using log-odds. The coefficient θ is interpreted as that term's contribution to the log-odds of an individual tie, conditional on all other dyads remaining the same.

Suppose θ represents the parameter for a triangle statistic. The log-odds interpretation would tell us how the presence of triangles in the network compares to networks without triangles (reference configuration). A positive estimate indicates a higher likelihood for that network configuration.

Modeling network data

Summing up with a social network example

Erdos-Renyi model Modeling random friendships where each pair of individuals has a small probability of being friend (usually used as baseline model)

Scale-free model Common in social media or citation networks, where a few nodes (influencers) dominate the network's structure

Small-world model Explaining the phenomenon where individuals are connected through a small number of intermediaries

PA model Modeling the growth of citation networks where papers with more citations attract more future citations (emergence of hubs)

ERG model Analyzing social networks to understand how network structure emerges from individual-level interactions in a regression fashion

SB model Identifying cohesive communities within social networks based on patterns of interactions or similarities in attributes

LS model Exploring hidden structures such as underlying preferences or affiliations that drive connection formation