From Egocentric Networks to Diffusion Processes: A Data-Driven Approach to Social Network Reconstruction

Thesis presentation

Stefano Di Filippo

Bocconi University

April 5, 2025

Table of contents

Motivation

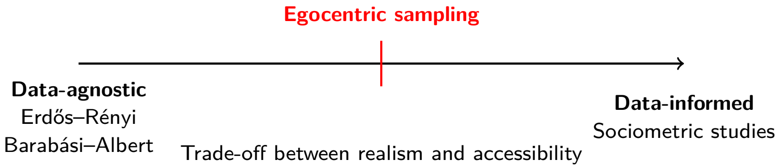
Understanding social dynamical processes requires the specification of an underlying network topology. This can be approached through:

- **Sociometric studies** — precise but costly.
- **Random graph models** (e.g., Erdős–Rényi, Barabási–Albert) — convenient but unrealistic.
- **Egocentric sampling** — accessible but limited to local information.

Motivation

Understanding social dynamical processes requires the specification of an underlying network topology. This can be approached through:

- **Sociometric studies** — precise but costly.
- **Random graph models** (e.g., Erdős–Rényi, Barabási–Albert) — convenient but unrealistic.
- **Egocentric sampling** — accessible but limited to local information.



Motivation (cont.)

Challenge: Reconstructing global network structures from egocentric sampling data to inform the dynamics of spreading processes in social systems.

Proposed Approach:

- Data-driven reconstruction using general egocentric information.
- Comparison with generative models (Erdős–Rényi, Barabási–Albert).

Goals:

- Assess differences in global graph properties.
- Evaluate impact on dynamic process evolution.

Contribution: Provide empirical evidence to guide future research in egocentric network reconstruction and dynamical modeling.

Egocentric Sampling

Egocentric Network Sampling

A sampling scheme in which a random sample of individuals (called *egos*) is surveyed about:

- their own characteristics,
- their social contacts (called *alters*),
- and possibly the ties among those alters.

The result is a collection of local networks centered on each ego, where nodes represent people and edges represent social ties.

Minimal design:

- Alters are not uniquely identified;
- Ties between alters are not collected.

Egocentric Network Analysis

ENA Goals:

- Explain and predict ego attributes based on their social environment;
- Describe the internal structure of ego-centered networks.

Two analytical dimensions:

- **Composition:** distribution and diversity of alter attributes
  (e.g., Blau's index, Shannon entropy, homophily measures like the E–I index);
- **Structure:** configuration of ties among alters
  (e.g., density, fragmentation index).

Egocentric Network Analysis

ENA Goals:

- Explain and predict ego attributes based on their social environment;
- Describe the internal structure of ego-centered networks.

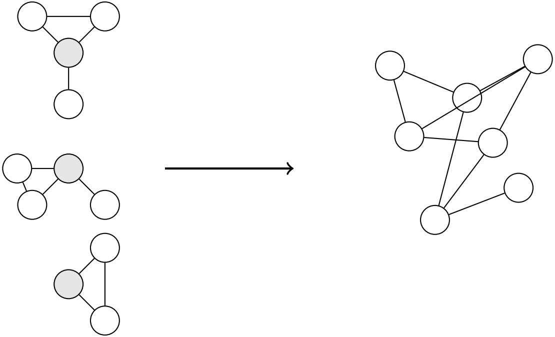
Two analytical dimensions:

- **Composition:** distribution and diversity of alter attributes (e.g., Blau's index, Shannon entropy, homophily measures like the E–I index);
- **Structure:** configuration of ties among alters (e.g., density, fragmentation index).

Limitation: ENA is inherently local—understanding global network topology and dynamics requires a model.

Next step: Reconstructing the global network structure from egocentric data.

From Local Views to Global Structure



Egocentric networks provide partial views of the whole.
We use models to reconstruct the global topology.

From local subnetworks... toward a generative reconstruction of the full system.

Modeling Approaches for Network Reconstruction

To reconstruct a global network from egocentric data, several classes of models can be considered:

- **Configuration Model**
  Generates random graphs constrained by node degrees.
  Captures degree distribution, but lacks clustering and higher-order structure.
- **Stochastic Block Models (SBMs)**
  Assign tie probabilities based on latent community membership.
  Particularly suited for networks with clear modular or community structure, but less commonly used for egocentric data reconstruction.
- **Exponential Random Graph Models (ERGMs)**
  Capture local structural patterns and attribute-based tendencies.
  Widely used in the literature and supported by existing tools for egocentric data.

Modeling Approaches for Network Reconstruction

To reconstruct a global network from egocentric data, several classes of models can be considered:

- **Configuration Model**
  Generates random graphs constrained by node degrees.
  Captures degree distribution, but lacks clustering and higher-order structure.
- **Stochastic Block Models (SBMs)**
  Assign tie probabilities based on latent community membership.
  Particularly suited for networks with clear modular or community structure, but less commonly used for egocentric data reconstruction.
- **Exponential Random Graph Models (ERGMs)**
  Capture local structural patterns and attribute-based tendencies.
  Widely used in the literature and supported by existing tools for egocentric data.

In this study, we focus on ERGMs due to their modeling flexibility, interpretability, and strong methodological foundation in the context of partial network data.

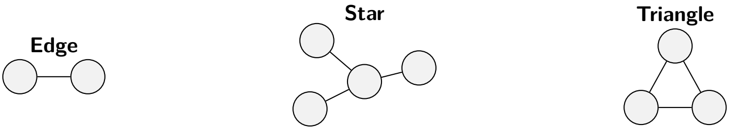
What Are Exponential Random Graph Models?

Exponential Random Graph Models (ERGMs) define a probability distribution over networks, where tie formation is modeled as a function of local structural patterns and node attributes.

Core Idea

The likelihood of a graph depends on the presence and frequency of specific network configurations.

Basic Configurations:



Other configurations include reciprocated ties, shared partners, and homophily effects.

Formal Specification of ERGMs

Let Y be a random network on n nodes and y its realization.

Model Form

$$P(Y = y|\theta) = \frac{\exp\left(\theta^T z(y)\right)}{K(\theta)}$$

Components:

- $z(y)$: vector of sufficient statistics (e.g., edge count, triangle count)
- θ: parameter vector controlling configuration importance
- $K(\theta)$: intractable normalizing constant

Why ERGMs for Egocentric Data?

Advantages in the Egocentric Context

- Can incorporate the statistics observable in ego-networks
- Compatible with partially missing or sampled data
- Flexible and interpretable modeling of social structure
- Supported by well-developed estimation frameworks and software

In this study, ERGMs allow us to translate egocentric observations into simulated global network topologies.

Modeling Transitivity and Avoiding Degeneracy

Problem: Simple ERGMs (e.g., with triangle terms) can produce unrealistic graphs due to *degeneracy*.

Solution: Geometrically Weighted Statistics (Snijders, 2006)

- **GWESP**: Geometrically weighted edgewise shared partners (captures transitivity)
- **GWD**: Geometrically weighted degree
- **GWDSP**: Geometrically weighted dyadwise shared partners

These statistics improve model stability and realism by damping the contribution of large counts.

These terms are essential for reproducing empirical social structures in simulated networks.

Estimation and Simulation of ERGMs

Challenge: The normalizing constant $K(\theta)$ involves summing over all possible networks.

Solution: Markov Chain Monte Carlo Maximum Likelihood Estimation (MCMC-MLE)

- Approximate the likelihood using MCMC sampling from the ERGM
- Iteratively update θ to align simulated and observed network statistics
- Common initialization: Maximum Pseudo-Likelihood Estimation (MPLE)

Simulation

- Networks are generated via Metropolis–Hastings steps (edge toggling)
- After convergence, the model can be used for simulation or evaluation

Evaluating ERGMs

Fitting an ERGM is not enough: we must assess how well it reproduces the data.

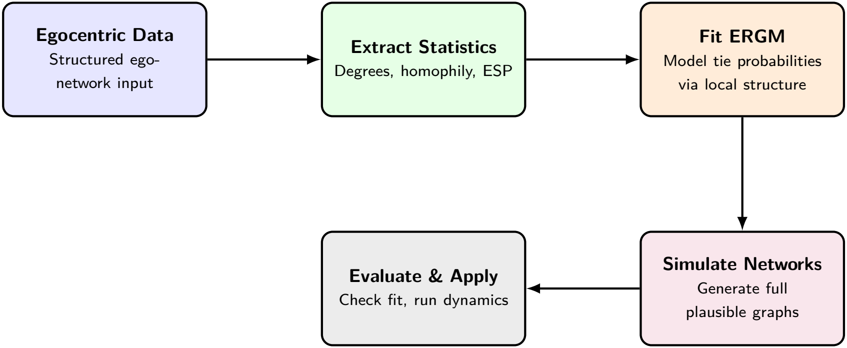
Goodness-of-Fit (GoF) in the Egocentric Context

- Simulate full networks from the fitted ERGM
- Compare simulated vs. observed values of statistics *extractable from egocentric data*
- Typical diagnostics: **degree distribution**, **clustering (local density)**, **edgewise shared partners (ESP)**, and other ego-level features

Note: Global metrics like geodesic distances cannot be validated against purely egocentric samples.

Goal: Ensure that the model accurately reproduces local structure—*a crucial step before using the simulated networks to study spreading processes.*

From Local Data to Simulated Global Networks



Goal: Use a data-informed model to reconstruct global networks and reproduce social processes.

From Methodology to Application

Transitioning from theory to empirical analysis

We now move from the conceptual framework to its implementation on real-world data.

Next: We apply the ERGM-based reconstruction pipeline to egocentric datasets, assessing its performance against benchmark models and examining the impact of reconstructed structure on diffusion dynamics.

Sampling Design in This Study

Our egocentric sampling scheme adopts a *partially minimal* design:

- **Alters are not uniquely identified**
  (i.e., we cannot detect overlaps across ego-networks);
- **Ties between alters are collected through proxies**
  (i.e., ego-reported measures of density, rather than direct alter–alter data).

This design is common in population-representative, topic-specific surveys:
it balances feasibility and informativeness by reducing respondent burden while preserving structural insight.

Egocentric Dataset: Structure and Scope

Source: ERC IMMUNE project (Melegaro), survey conducted by Bocconi University on health-related decision-making and social influence.

Sample:

- $n = 12{,}122$ respondents (egos) from 4 countries: Germany, France, Italy, UK
- Focus on two distinct social layers: **friends** and **coworkers**

Collected Information:

- **Egos:** sex, age group, education, income, political orientation, household composition
- **Alters:** age group only
- **Network structure:**
  - Ego–alter ties (count-based, separately for friends and coworkers)
  - Alter–alter ties: reported via categorical density proxy (e.g., low/medium/high)

Objective: Reconstruct global networks of friendship and coworking ties, separately by country.

Preprocessing Steps for Network Reconstruction

Cleaning and Filtering:

- Removed incomplete or implausible records (e.g., extremely high alter counts)
- Filtered out alters under 18 years old (no minors among egos)

Handling Alter–Alter Ties:

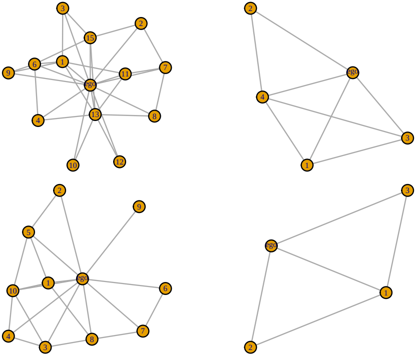
- Categorical density proxy mapped to link probabilities: *(e.g., "medium" → 0.5 density)*
- Edges assigned at random within ego-networks accordingly

Smoothing Degree Distributions:

- Observed rounding bias (multiples of 5) in ego-reported degrees
- Applied a two-step correction:
  - Binned and redistributed degrees uniformly within local windows
  - Fitted a Gamma distribution to smooth and re-sample realistic degree values

Outcome: Cleaned, smoothed ego-network dataset ready for ERGM fitting.

Examples of Observed Egocentric Networks



Each subgraph illustrates a respondent's social environment:

- The central node is the **ego**.
- Peripheral nodes are **alters** (friends or coworkers).
- Edges among alters are inferred via ego-reported density proxies.

These networks vary in size and internal cohesion, highlighting the diversity of ego-centered structures in our dataset.

ERGM Specifications for Network Reconstruction

Separate models for:

- Coworker networks
- Friendship networks

Included terms:

- **Node covariates:** sex, age group, education, income, political orientation, household structure
- **Differential homophily** on nodal attributes
- **GWD** (degree heterogeneity) and **GWESP** (transitivity)

Decay parameters:

- Fixed: (0.25, 0.25), (0.5, 0.5), (0.8, 0.8), (1, 0.2), (1, 0.5), GWD only, GWESP only
- Free: estimated during fitting

Estimation Procedure

Method: Geyer–Thompson MCMC-MLE

Main settings:

- **Iterations:** 50
- **Samples per chain:** 2000 (interval: 2000 steps)
- **Burn-in:** 2000 steps
- **Constraint:** Max degree fixed to empirical max

Software: `statnet` R package suite

We also test unconstrained runs with lighter MCMC settings.

Model Evaluation: Goodness of Fit

Approach: simulate 100 networks from each ERGM and compare with observed data.

Metrics used:

- Degree distribution
- Edgewise shared partner (ESP) distribution

Evaluation method: Kullback–Leibler divergence to assess closeness of empirical and simulated distributions

Note: global metrics like distances are not suitable due to egocentric design.

Benchmark Models for Reconstruction

Final step: Use best ERGMs to generate global networks for:

- Coworkers
- Friends

Compared with:

- **Erdős–Rényi graph:** Poisson random graph with empirical density
- **Barabási–Albert model:** Scale-free network with $m = \langle k \rangle / 2$

Main comparison metrics:

- Degree distribution
- Average shortest path
- Path length distribution
- Clustering coefficient

Simulating Dynamical Processes

Goal: Assess the structural implications of network reconstruction by simulating two spreading processes:

- **Epidemic Spreading (SEIR model)** — disease-like dynamics
- **Opinion Dynamics (Friedkin–Johnsen model)** — social influence

Both models are applied to the reconstructed networks of friends and coworkers (after isolating the largest connected component).

SEIR Model of Epidemic Spreading

States: Susceptible (S), Exposed (E), Infectious (I), Recovered (R)

$$\mathbf{S} \rightarrow \mathbf{E} \rightarrow \mathbf{I} \rightarrow \mathbf{R}$$

Transition rules:

- S $\rightarrow$ I: exponential rate β
- E $\rightarrow$ I: gamma-distributed waiting time (k_E, θ_E)
- I $\rightarrow$ R: gamma-distributed waiting time (k_I, θ_I)

Simulations: 45 parameter combinations using the `epinet` R package

SEIR Simulation Parameters

Parameter	Values
Transmission Rate (β)	0.2, 0.3, 0.4, 0.5, 0.6
Incubation Rate (σ)	1/3, 1/4, 1/5
Recovery Rate (γ)	1/5, 1/6, 1/7
Total Simulations	45

Note: The variance of incubation and infection periods is set at 0.5.

Friedkin–Johnsen Model of Opinion Dynamics

Opinion evolution:

$$p_i(t+1) = \frac{(1-\beta)\, p_i(0) + \beta \sum_{j \in N(i)} p_j(t)}{1-\beta+\beta k_i}$$

Assumptions:

- $p_i(0)$ is the internal opinion (fixed)
- $p_i(t) \in [0,1]$ at all times
- $\beta \in [0,1]$: strength of social influence

Simulation ends when opinions converge.

Friedkin–Johnsen Simulation Parameters

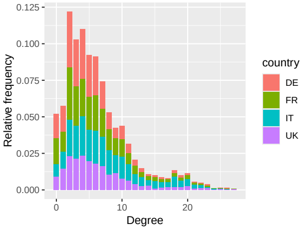
Parameter	**Values**
Influence Effect (β)	$\{0.05, 0.10, \dots, 1.00\}$
Internal Opinion Shape 1 (α_1)	0.1, 0.5, 1
Internal Opinion Shape 2 (α_2)	0.1, 0.5, 1
Total Simulations	180

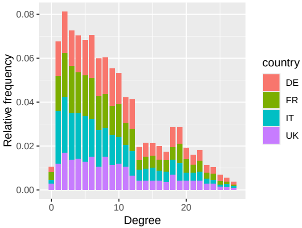
Note: Internal opinions $p_i(0)$ drawn from $\text{Beta}(\alpha_1, \alpha_2)$.

Exploratory Analysis: Degree Distributions and Age Effects

Goal: Investigate structural differences between friends and coworkers networks, assess attribute effects, and inform modeling.

- **Friends Network:** More degree variability, with a higher concentration of low-degree nodes.
- **Coworkers Network:** Degree distribution is more concentrated around the average.

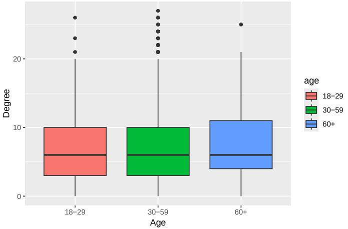




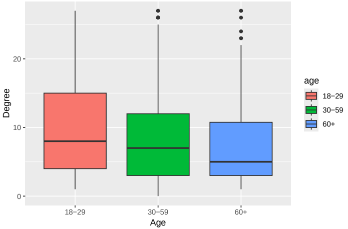
Left: Friends Network. Right: Coworkers Network.

Exploratory Analysis: Age-Specific Degree Distributions

Question: Is age associated with degree across ego networks?

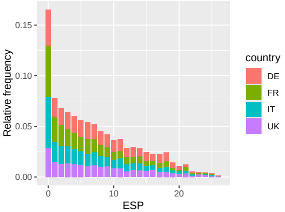


- Friends: no strong age pattern
- Coworkers: some variation by age group
- Suggests age may influence coworker connectivity more than friendships

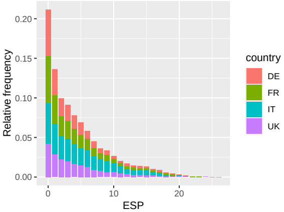


Top: Friends. Bottom: Coworkers.

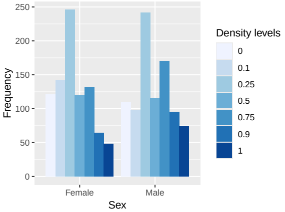
Transitivity and Local Density Patterns



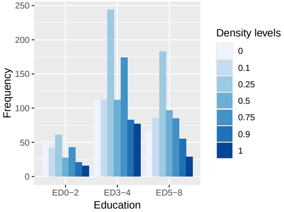
ESP Distribution (Coworkers)



ESP Distribution (Friends)



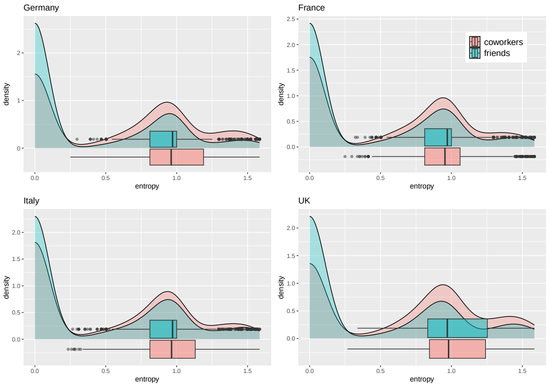
Friends by Sex (Italy)



Friends by Education (Italy)

Insight: Friends show lower ESP values → less transitivity. Local density varies by education, less so by gender.

Age Diversity Within Ego Networks



Density: Shannon entropy of age group diversity across ego networks.
Boxplot: Distribution of entropy for egos with non-zero diversity.

Insight: Friends networks are more often age-homogeneous.
Coworkers show higher variability in alter age diversity.

Age Homophily Patterns: Mixing Matrices by Relationship Type

Coworkers Network

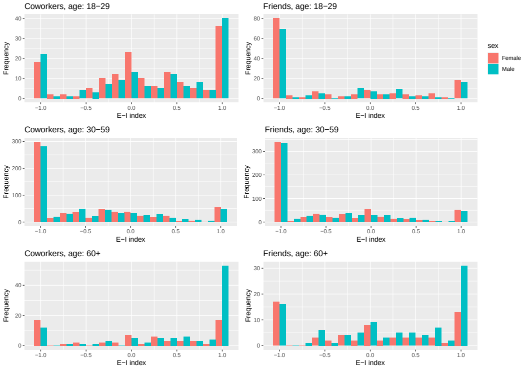
	Germany			France		
	18–29	30–59	60+	18–29	30–59	60+
18–29	0.09	0.10	0.02	0.10	0.11	0.01
30–59	0.13	0.49	0.06	0.16	0.51	0.05
60+	0.02	0.08	0.02	0.01	0.05	0.01
	Italy			UK		
	18–29	30–59	60+	18–29	30–59	60+
18–29	0.09	0.10	0.02	0.14	0.11	0.02
30–59	0.12	0.55	0.08	0.14	0.35	0.05
60+	0.01	0.06	0.02	0.02	0.06	0.02

Friends Network

	Germany			France		
	18–29	30–59	60+	18–29	30–59	60+
18–29	0.15	0.05	0.01	0.15	0.04	0.01
30–59	0.10	0.51	0.06	0.12	0.53	0.08
60+	0.01	0.05	0.05	0.01	0.04	0.03
	Italy			UK		
	18–29	30–59	60+	18–29	30–59	60+
18–29	0.11	0.04	0.01	0.20	0.06	0.02
30–59	0.10	0.57	0.08	0.11	0.43	0.06
60+	0.01	0.05	0.04	0.01	0.05	0.05

Insight: Across countries, both networks show clear age homophily. Friends networks exhibit even stronger within-group connections among younger individuals.

E-I Index for Age Attribute (Italian Egos)



E-I Index: Measures the tendency of egos to connect to alters **External** to vs **Internal** to a category (e.g., age group).

Interpretation:

- Negative values = homophily (within-group preference)
- Positive values = heterophily (cross-group ties)

Findings:

- Strong homophily for middle-aged egos.
- Heterophily appears among young coworkers.
- Gender differences emerge mostly among older individuals.

Model Estimation for the Friends Network

Selected ERGM Specification:

- **Decay Parameters:** GWD (Decay = 1), GWESP (Decay = 0.8)
- **Significant Effects:**
  - **Age Homophily:** Highly significant for all age groups
  - **Income:** Significant effect on social ties (e.g., Quintile 2 shows weak negative effect)
- **Model Fit:** KL Divergence for degree distribution: ~0.05

Transitivity: GWESP term is significant but does not fully capture ESP distribution due to sample space constraints.

Next: Compare fitted and empirical degree and ESP distributions.

Model Estimation for the Coworkers Network

Selected ERGM Specification:

- **Decay Parameters:** GWD (Decay = 0.8), GWESP (Decay = 0.8)
- **Significant Effects:**
  - **Age Homophily:** Significant for all age categories
  - **Education:** Significant effect on coworker ties (e.g., Secondary Education: positive effect)
- **Model Fit:** KL Divergence for degree distribution: ~0.16

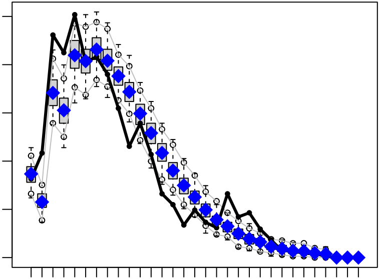
Transitivity: GWESP term is positive but does not fully capture ESP distribution, similar to the friends network.

Next: Compare fitted and empirical degree and ESP distributions.

Model Estimation - Transitivity and Degree Fit

Comparing Empirical and Simulated Distributions:

- **Degree Distribution:** Simulated degree distribution closely matches the empirical data, with minor discrepancies.
- **ESP Distribution:** While the GWESP term captures some transitivity, the simulation shows a significant difference from the empirical distribution.



Note: Further evaluation through KL divergence and statistical comparison will follow.

Model Estimation – Improving Fit

Improvement in KL Divergence After Relaxing Constraints:

- **Friends Network:** KL divergence decreases from 4.17 to 3.89 after relaxing sample space constraints.
- **Coworkers Network:** KL divergence decreases from 3.52 to 2.74 after relaxing sample space constraints.

Transitivity Challenges:

- Despite the improvement in degree and edgewise shared partner distributions, transitivity is still not fully captured.
- Further refinements are needed to achieve a more accurate representation of the observed transitivity patterns.

Note: These results indicate progress, but the model still faces challenges in fully replicating the network's transitive structure.

Model Evaluation – Overview

Comparing Network Topologies:

- Evaluating the ERGM networks against Erdős–Rényi (ER) and Barabási-Albert (BA) models.
- Both global network measures and the evolution of spreading processes are compared.
- Simulations are run for 10 graphs of 10,000 nodes for each network topology.

Next: Detailed comparison using key metrics.

Model Evaluation – Coworkers Network Metrics

Metrics Comparison for Coworkers Network:

- Comparison of key metrics: Average degree, degree standard deviation, average distance, clustering, and centrality measures.
- The ERGM shows superior performance in capturing both the degree distribution and transitivity.

Coworkers Network Metrics:

Metric	ER	BA	ERGM
Average Degree	8.38	8.00	8.45
Degree SD	2.90	9.14	6.11
Average Distance	4.57	4.02	4.70
Distance SD	0.19	0.29	0.39
Global Clustering	< 0.01	< 0.01	0.20
Average Clustering	< 0.01	< 0.01	0.36
Average Closeness Centrality	2.19e-05	2.50e-05	2.17e-05
Closeness Centralization	0.06	0.27	0.13
Average Betweenness Centrality	1.78e-04	1.51e-04	1.83e-04
Betweenness Centralization	0.002	0.08	0.01
Average Eigenvalue Centrality	0.28	0.13	0.14
Eigenvalue Centralization	0.72	0.99	0.99

Table 1: Metrics of reconstructed coworkers networks.

Model Evaluation – Friends Network Metrics

Metrics Comparison for Friends Network:

- Comparison of key metrics: Average degree, degree standard deviation, average distance, clustering, and centrality measures.
- The ERGM shows superior performance in capturing both the degree distribution and transitivity.

Friends Network Metrics:

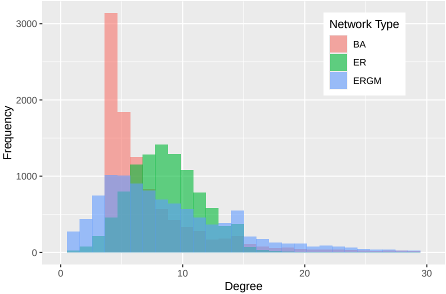
Metric	ER	BA	ERGM
Average Degree	7.61	6.00	7.72
Degree SD	2.77	6.86	5.55
Average Distance	4.77	4.48	4.92
Distance SD	0.22	0.36	0.41
Global Clustering	< 0.01	< 0.01	0.21
Average Clustering	< 0.01	< 0.01	0.37
Average Closeness Centrality	2.19e-05	2.50e-05	2.17e-05
Closeness Centralization	0.05	0.26	0.12
Average Betweenness Centrality	1.88e-04	1.74e-04	1.90e-04
Betweenness Centralization	0.002	0.12	0.01
Average Eigenvalue Centrality	0.32	0.01	0.13
Eigenvalue Centralization	0.68	0.99	0.99

Table 2: Metrics of reconstructed friends networks.

Model Evaluation – Comparison of Key Metrics

Degree Distribution :

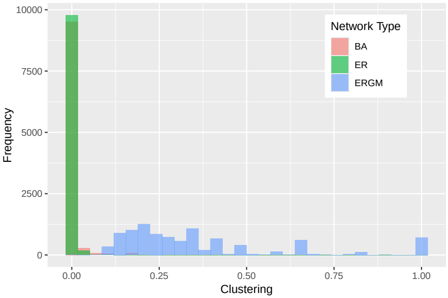
- ERGM closely matches the degree distribution seen in the empirical data.
- ERGM exhibits more nodes of very low degree.



Model Evaluation – Distribution of Local Clustering Coefficients

Local Clustering Coefficients:

- ERGM exhibits a broad distribution of local clustering coefficients, from 0 to 1.
- In contrast, ER and BA models show clustering concentrated near 0.

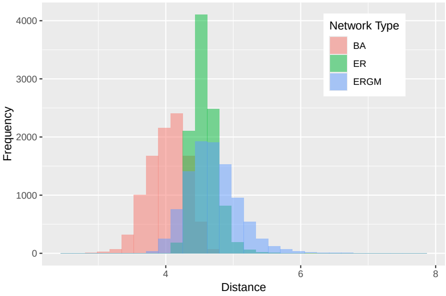


Note: ERGM provides more realistic clustering patterns.

Model Evaluation – Distribution of Network Distances

Average Distance:

- ERGM networks exhibit an average distance of 4.77, indicative of the small-world effect.
- Similar to BA but with greater variability compared to ER models.

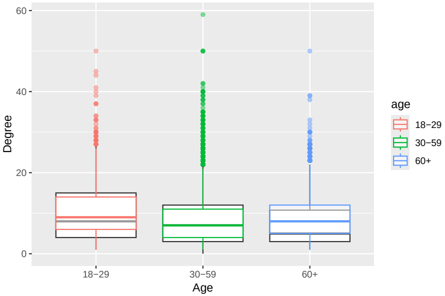


Note: Distances suggest realistic network structure across models.

Model Evaluation – Degree Distribution by Age Group (Coworkers)

Degree Distribution by Age:

- ERGM captures the age-related effects present in the empirical data.
- The degree distribution by age shows reasonable alignment between the empirical data and the ERGM.

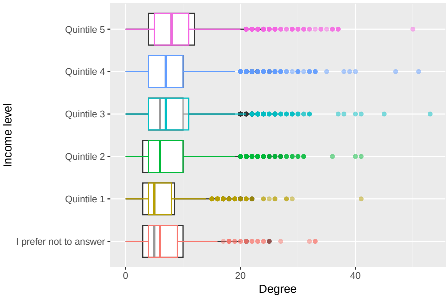


Note: This highlights the ability of ERGM to reflect demographic patterns.

Model Evaluation – Degree Distribution by Income (Friends)

Degree Distribution by Income:

- ERGM reconstructs the degree distribution with respect to income levels.
- The reconstructed network captures the varying impact of income on friendships.

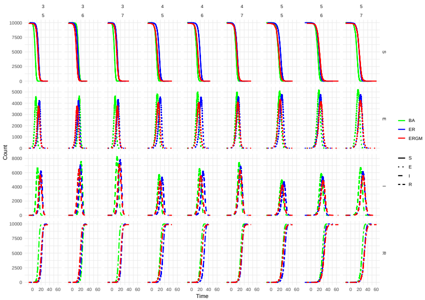


Note: ERGM captures the relationship between income and degree distribution well.

Model Evaluation – SEIR Epidemic Simulation

SEIR Epidemic Simulation:

- ERGM networks show a slower epidemic spread compared to ER and BA networks.
- The maximum number of infectious individuals is smaller in the ERGM network.



Note: **ERGM** is shown in red, **ER** in blue, and **BA** in green. The ERGM network slows the epidemic spread across a range of SEIR parameters.

Opinion Dynamics Model – Final Opinions (Influence Parameter 0.1 and 0.5)

Overview of the Final Opinion Dynamics:

- Simulations across all influence parameters show differences in the final opinion distributions. In the ERGM more extreme positions survive.
- The plots show the difference across varying influence parameters and initial opinion distributions.

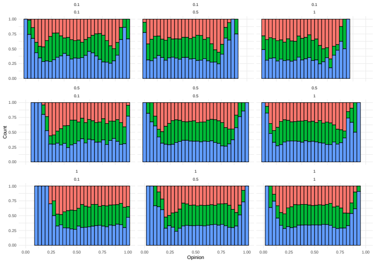


Figure 1: Influence parameter 0.1

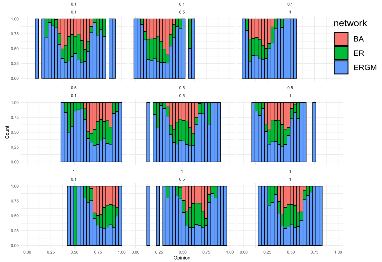


Figure 2: Influence parameter 0.5

Opinion Dynamics Model (cont.)

Overview of the Opinion Dynamics for Different Influence Parameters:

- Simulations across all influence parameters show differences in the final opinion distributions. In the ERGM more extreme positions survive.
- The plots show the difference across varying influence parameters and initial opinion distributions.

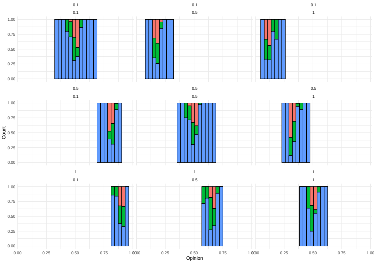


Figure 3: Influence parameter 0.9

Opinion Dynamics Model (cont.)

Opinion Dynamics:

- ERGM networks exhibit more polarization in final opinions compared to other networks.
- Differences in final opinions are evident even at low influence levels.

Note: The plot shows the range of final opinions across all simulations. The ERGM network maintains more extreme opinions.

Conclusion – Study Overview

Key Findings:

- The study demonstrates the potential of ego-centered data for reconstructing networks and simulating dynamical systems.
- The ERGM-based reconstruction method successfully captured main features of the data and distinctive patterns in global metrics.
- Despite simplifications in ERGM fitting, the results provided valuable insights into network dynamics.
- Future work should focus on overcoming computational challenges and refining methods to make ego-based network modeling more efficient and scalable.
- The integration of ego-centered data will bridge the gap between network and social sciences, enhancing precision in modeling and simulations.

Code Repository

You can find the code for this thesis at:
https://github.com/stefano-dif/thesis_egonet-ERGM