Ca' Foscari
University
of Venice

Master Degree
in Data Analytics for Business and Society

Final Thesis

Reconstructing Interbank Networks for Counterfactual Policy Evaluation: A GNN-IPF Hybrid Approach with Application to Basel III Capital Buffers

Supervisor
Ch. Prof. Giulia Iori

Graduand
Ali Ebadatian
Matriculation Number 903878

Academic Year
2025 / 202

To Niloofar.

Chapter 1

Introduction

The stability of the financial system depends critically on the network of bilateral exposures linking banks to one another. When a bank defaults, losses propagate through these interbank linkages to its counterparties, potentially triggering cascading failures that amplify an initially localized shock into a systemic crisis. The global financial crisis of 2007–2009 demonstrated that neither regulators nor market participants possessed an adequate map of these interconnections, a deficiency that contributed to the mispricing of counterparty risk and the underestimation of systemic fragility (Acemoglu et al., 2015; Eisenberg and Noe, 2001).

A fundamental obstacle to monitoring interbank contagion risk is that the complete matrix of bilateral exposures is almost never observed. Central banks and supervisory authorities typically have access only to aggregate balance sheet positions — total interbank assets and liabilities for each institution — reported through regulatory filings such as those compiled by the Banca d'Italia or extracted from commercial databases like Bureau van Dijk's ORBIS. The individual bilateral claims that compose these aggregates remain confidential. This information asymmetry creates what Anand et al. (2018) term the "missing links" problem: regulators must assess systemic risk on a network they cannot directly observe, armed only with marginal totals that constrain the row and column sums of an unknown matrix.

The standard response in the literature has been to reconstruct the bilateral exposure matrix from these marginal constraints using statistical or information-theoretic methods. The maximum entropy approach, introduced to interbank networks by Upper and Worms (2004) and Sheldon and Maurer (1998), selects the matrix that maximizes Shannon entropy subject to the observed marginals, producing a dense, nearly complete network in which exposures are spread as uniformly as possible. While analytically tractable and widely adopted by central banks, this method systematically underestimates concentration risk by dispersing exposures across too many counterparties (Anand et al., 2015; Mistrulli, 2011). At the opposite extreme, the minimum density method of Anand et al. (2015) seeks the sparsest matrix consistent with the marginals, producing a concentrated but often unrealistically sparse topology. The fitness model of Cimini et al. (2015) occupies a middle ground, assigning connection probabilities proportional to banks' interbank activity, but remains purely size-driven and cannot capture structural features such as reciprocity, disassortativity, or community structure that empirical interbank networks are known to exhibit (Iori et al., 2008; Precup et al., 2005; Fricke and Lux, 2015).

The core limitation shared by all these classical approaches is that they are *memoryless*: they reconstruct each period's network from scratch using only contemporaneous marginals, discarding any information that might be learned from the structural regularities of past networks. Yet interbank markets exhibit persistent topological features — stable core-periphery

structures, characteristic degree distributions, and recurring patterns of reciprocal trading — that carry information about which bilateral links are likely to exist in a given period. A reconstruction method that could learn these patterns from historical data and incorporate them as informed priors should, in principle, outperform methods that treat every period as a blank slate.

This thesis proposes such a method. We develop a hybrid pipeline that combines a Graph Neural Network (GNN) trained on historical interbank transaction data with Iterative Proportional Fitting (IPF) to reconstruct weighted directed interbank networks from marginal information alone. The GNN, a two-layer encoder based on the TransformerConv architecture of Shi et al. (2021), is trained on daily transaction graphs from the Italian e-MID interbank platform over the period 2000–2005. The model learns to predict edge existence from node-level features — lagged annual volumes, institutional group membership, and macroeconomic variables — using a gravity-based scaffold that provides the encoder with a structurally plausible initial topology without exposing it to the true edges it is tasked with predicting. A multi-objective loss function jointly optimizes link prediction accuracy, reciprocity matching, degree distribution fidelity, and disassortative mixing, with weights selected through a grid search over 15 configurations targeting the specific microstructural properties identified by Precup et al. (2005) as characteristic of the Italian overnight market.

The trained GNN produces a matrix of edge probabilities that serves as an informed prior for the second stage of the pipeline. The top-k most probable edges are selected and weighted using a gravity kernel derived from the observed marginals, then passed as a seed matrix to the IPF algorithm, which iteratively rescales the matrix until it exactly satisfies the row and column sum constraints. This two-stage design exploits the complementary strengths of each component: the GNN contributes topological realism — it knows which edges are structurally plausible given the market's historical patterns — while IPF ensures volumetric consistency with the observed balance sheet data.

We evaluate the pipeline on the out-of-sample reconstruction of the Italian interbank network for the period 2006–2007, where the true bilateral transaction matrix from e-MID serves as ground truth. The GNN+IPF method is compared against three established baselines — maximum entropy, minimum density, and the fitness model — across multiple dimensions: edge prediction accuracy, microstructural fidelity (degree distribution, clustering, strength-degree correlation, disassortativity), and centrality preservation. The fitness model achieves marginally higher F1 scores on edge identification (0.692 versus 0.674), and both methods perform comparably on reciprocity and assortativity at the bi-annual aggregation level. However, GNN+IPF significantly outperforms all baselines on centrality preservation, with a mean Spearman rank correlation of 0.926 across eight centrality measures versus 0.905 for fitness (Wilcoxon $p = 0.004$). The advantage is concentrated in authority score ($+0.071$ over fitness) and betweenness centrality ($+0.046$), the measures most relevant for identifying contagion intermediaries. GNN+IPF also achieves an exact match on the Gini coefficient of out-strength — a metric absent from the training objective — while the fitness model deviates by 0.023. An ablation study confirms that these advantages are attributable to the GNN encoder rather than the gravity scaffold alone: the scaffold achieves comparable edge-level accuracy but fails to match the Gini coefficient (error 0.088 versus 0.000) or preserve PageRank rankings ($\rho = 0.891$ versus 0.983).

To demonstrate the policy relevance of improved reconstruction, we apply the pipeline to a counterfactual analysis of Basel III macroprudential capital buffers on the contemporary Italian banking system. Using balance sheet data from ORBIS for a common set of 188 Italian commercial banks present in both the 2023 and 2024 filings, we adopt a temporal design that

mirrors operational practice: systemic risk scores are computed on the 2023 reconstructed network, buffer surcharges are frozen, and the resulting capital requirements are applied to the 2024 network for stress testing. This out-of-sample approach avoids the in-sample circularity of scoring and testing on the same network. Capital buffers are assigned through three scoring mechanisms — EBA indicator-based scoring, DebtRank impact scoring, and DebtRank vulnerability scoring — and implemented via a deleveraging channel in which undercapitalized banks reduce interbank lending. The analysis reveals that score-based buffer regimes reduce aggregate systemic risk by approximately 12% relative to the unregulated baseline, with all three scoring mechanisms producing similar outcomes (11.9–12.7% risk reduction). The temporal validation across 30 Monte Carlo seeds confirms that the optimism gap between in-sample and out-of-sample scoring is below 0.05 percentage points, and the set of deleveraging banks is identical across years (Jaccard = 1.000).

A notable finding is that the choice of reconstruction method does not materially affect the policy conclusions in this setting. The GNN+IPF and fitness reconstructions produce nearly identical systemic risk estimates on the 2024 network (aggregate DebtRank of 0.554 versus 0.552), despite their different topological properties. This convergence occurs because the contagion dynamics in the well-capitalized Italian banking system are dominated by the *volume* of bilateral exposures — which IPF equalizes across methods — rather than their *pattern*. The structural advantages of GNN+IPF documented in the reconstruction evaluation improve the fidelity of the network representation but do not change the direction or magnitude of the policy recommendations in the small-shock regime considered. Whether this convergence holds under larger shocks or in systems with binding capital constraints remains an open question.

The robustness of these findings is established through seven tests: a 30-seed Monte Carlo over random initializations, an edge-budget sensitivity sweep, a temporal buffer experiment on e-MID ground truth data, a distributional shift analysis quantifying the e-MID-to-ORBIS domain gap, LGD sensitivity analysis, four alternative feature-mapping specifications (baseline, relative exposure, rank-based, and hybrid), and a sub-period training analysis across four temporal windows. All qualitative conclusions survive these perturbations.

The remainder of this thesis is organized as follows. Chapter 2 reviews the literature on interbank networks, reconstruction methods, graph neural networks, and contagion modeling. Chapter 3 describes the data sources: the e-MID platform, macroeconomic indicators, and the ORBIS database. Chapter 4 presents the methodology in detail, covering the GNN architecture, the IPF integration, the baseline methods, and the DebtRank framework. Chapter 5 reports the reconstruction results, including microstructure analysis and centrality preservation tests. Chapter 6 applies the pipeline to the contemporary Italian banking system and evaluates capital buffer scenarios using the temporal design. Chapter 7 presents the robustness analysis. Chapter 8 concludes with a discussion of limitations and directions for future research.

Chapter 2

Literature Review

This chapter reviews the four bodies of literature that converge in this thesis: the theory of systemic risk in financial networks, the methods developed to reconstruct bilateral exposures from partial data, the application of graph neural networks to network inference problems, and the modeling of contagion dynamics for macroprudential policy evaluation. We begin with the theoretical foundations that motivate the problem, proceed through the classical reconstruction methods that define our baselines, survey the machine learning approaches that inspire our methodology, and conclude with the contagion framework that provides the policy application.

2.1 Interbank Networks and Systemic Risk

The modern analysis of systemic risk through the lens of network theory begins with the observation that the pattern of bilateral exposures among financial institutions — not merely their individual balance sheets — determines the system's vulnerability to cascading failures. Allen and Gale (2000) provide the foundational theoretical framework, showing that the structure of interbank linkages determines whether shocks are absorbed or amplified. In their model, complete networks where every bank is connected to every other bank facilitate risk sharing by distributing losses widely, while incomplete networks can concentrate losses on specific institutions and trigger contagion cascades.

Eisenberg and Noe (2001) formalize the clearing mechanism for interbank obligations, establishing the mathematical framework for computing equilibrium payment vectors in a network of interconnected banks. Their model treats the interbank liability matrix as given and solves for the set of payments that each bank can make given its assets and the payments it receives from others. The resulting fixed-point problem admits a unique clearing vector under mild conditions, and default cascades emerge endogenously when some banks' assets are insufficient to cover their obligations. This clearing framework has become the standard building block for computational models of interbank contagion.

Acemoglu et al. (2015) provide the most complete theoretical treatment of the relationship between network structure and systemic stability. They identify a fundamental tension: in the absence of large shocks, denser interbank networks are more stable because they distribute risk more evenly, consistent with Allen and Gale's intuition. However, beyond a critical shock threshold, this relationship reverses: dense networks become channels of contagion that propagate losses to otherwise healthy institutions, making the system more fragile. This phase transition result has profound implications for network reconstruction: methods that systematically produce denser networks (such as maximum entropy) will underestimate systemic risk in normal times but overestimate the network's absorptive capacity under stress, while meth-

ods that produce sparser networks (such as minimum density) may overestimate fragility under small shocks but better capture tail risk. The implication is that getting the density right — and more generally, getting the topology right — matters for the accuracy of systemic risk assessments.

Empirical work has confirmed that real interbank networks exhibit characteristic structural features that deviate markedly from both the complete and the random graph benchmarks. Iori et al. (2008) provide the definitive empirical study of the Italian interbank market using transaction-level data from the e-MID platform. They document a core–periphery structure in which a small set of highly connected banks intermediates between a large periphery of less active institutions. The degree distribution is heavy-tailed, consistent with a scale-free topology, and the network exhibits pronounced disassortativity: highly connected banks tend to trade with less connected banks rather than with each other. These features are not unique to Italy. Fricke and Lux (2015) confirm the core–periphery structure on e-MID using formal block-model methods and show that it is stable over time, while Degryse and Nguyen (2007) document similar patterns in the Belgian interbank market and Cont et al. (2013) in the Brazilian system. The universality of these structural regularities suggests that reconstruction methods should reproduce them if they are to be useful for systemic risk analysis.

Precup et al. (2005) provide the microstructural benchmarks that we adopt in our evaluation framework. Analyzing the Italian overnight market on e-MID, they characterize four key properties: the power-law exponent of the degree distribution, the strength–degree correlation (measuring whether well-connected banks also trade larger volumes), the mean clustering coefficient (measuring the tendency of a bank's counterparties to also trade with each other), and the nearest-neighbor degree function $k_{nn}(k)$ (measuring disassortative mixing). We use these four properties as the basis for our microstructural comparison between reconstruction methods and ground truth.

2.2 Network Reconstruction from Partial Data

The problem of reconstructing a bilateral exposure matrix from marginal information has generated a substantial literature, surveyed comprehensively by Upper (2011) and Anand et al. (2018). The setting is as follows. The regulator observes, for each bank i, the total interbank assets $a_i = \sum_j W_{ij}$ (row sums) and total interbank liabilities $l_j = \sum_i W_{ij}$ (column sums) of an unknown $n \times n$ matrix W of bilateral exposures, where $W_{ij} \geq 0$ represents the claim of bank i on bank j, with $W_{ii} = 0$. The problem is severely underdetermined: $n^2 - n$ unknowns are constrained by only $2n$ equations, so any reconstruction requires additional assumptions.

2.2.1 Maximum Entropy

The maximum entropy approach, introduced to the interbank context by Sheldon and Maurer (1998) and Upper and Worms (2004), selects the matrix that maximizes the entropy functional

$$H(W) = -\sum_{i \neq j} \frac{W_{ij}}{T} \ln \frac{W_{ij}}{T}, \qquad T = \sum_{i \neq j} W_{ij}, \tag{2.1}$$

subject to the marginal constraints $\sum_j W_{ij} = a_i$ and $\sum_i W_{ij} = l_j$. The solution distributes exposures as uniformly as possible across all potential counterparties, yielding a nearly complete matrix in which $W_{ij} = a_i l_j / T$ for all $i \neq j$. The solution can be computed directly or obtained as the limit of the IPF algorithm applied to a uniform seed matrix (Bacharach, 1965).

The maximum entropy method has been widely adopted by central banks for stress testing precisely because of its simplicity and unique solution. However, its empirical performance has been consistently criticized. Mistrulli (2011) provides the most direct evidence, comparing maximum entropy reconstructions against the true bilateral exposure matrix of the Italian banking system (obtained from the Banca d'Italia's Central Credit Register) and finding that the method systematically underestimates the severity of contagion. The mechanism is intuitive: by spreading exposures across all possible counterparties, maximum entropy eliminates the concentration of risk on specific bilateral links that drives real-world contagion. Upper (2011) reaches similar conclusions across multiple countries, noting that the method's performance deteriorates precisely in the scenarios of greatest regulatory interest — large shocks to major institutions.

From a topological perspective, the maximum entropy matrix is pathological: it produces a fully connected network with near-zero clustering variation, no meaningful degree heterogeneity, and assortativity coefficients indistinguishable from zero. These features are inconsistent with every empirical study of interbank network structure.

2.2.2 Minimum Density

Anand et al. (2015) propose an alternative that addresses the over-connectivity of maximum entropy. Their minimum density method seeks the sparsest matrix consistent with the marginal constraints, formulated as an optimization problem that minimizes the number of positive entries in W subject to the row and column sum constraints. In its exact form, this is an integer programming problem; practical implementations use LP relaxations or heuristic algorithms. The method produces networks with realistic sparsity levels but introduces a different distortion: it concentrates exposures on a minimal set of links, producing networks that are too sparse and too concentrated relative to empirical benchmarks. In our implementation, we follow the stochastic variant available in the `NetworkRiskMeasures` R package (Cinelli, 2017), which uses Gibbs-Boltzmann link probabilities with iterative edge removal to approximate the minimum density solution.

2.2.3 Fitness Model

The fitness model of Cimini et al. (2015) takes a different approach rooted in the statistical mechanics of complex networks. Rather than optimizing an objective function subject to constraints, it assigns each bank a pair of "fitness" parameters — one for lending propensity x_i and one for borrowing propensity y_j — calibrated to match the observed marginals. The probability of a directed link from i to j is then

$$p_{ij} = \frac{z x_i y_j}{1 + z x_i y_j}, \tag{2.2}$$

where z is a global parameter calibrated so that the expected number of links matches a target density. The fitness parameters are set to the observed marginals ($x_i = a_i$, $y_j = l_j$), making the model entirely determined by balance sheet data.

The fitness model represents a significant improvement over maximum entropy in that it produces heterogeneous networks with realistic density and degree distributions. Cimini et al. (2015) show that it outperforms maximum entropy on multiple reconstruction benchmarks. However, the model remains purely size-driven: link probabilities depend only on the product of marginals, not on any learned structural features. This means it cannot capture reciprocity

(the tendency for bank i to lend to bank j if j also lends to i), disassortative mixing (the tendency for large banks to connect to small banks rather than to each other), or any other topological pattern that is not a direct consequence of heterogeneous bank sizes.

Squartini et al. (2018) provide a comprehensive comparison of reconstruction methods within a unified statistical framework, showing that fitness-based approaches dominate entropy-based ones on most topological metrics but still fail to capture higher-order structural properties. Macchiati et al. (2022) extend the fitness framework by explicitly enforcing density and reciprocity constraints, demonstrating that conditioning on additional network statistics improves reconstruction quality. Their work motivates our approach of using a GNN to learn multiple structural features simultaneously, rather than imposing them as analytical constraints.

2.2.4 Iterative Proportional Fitting

Iterative Proportional Fitting (IPF), also known as the RAS algorithm or bi-proportional fitting, is the computational workhorse underlying most reconstruction methods. Given a non-negative seed matrix $W^{(0)}$ and target row sums $\mathbf{a}$ and column sums $\mathbf{l}$, IPF alternates between scaling rows to match $\mathbf{a}$ and scaling columns to match $\mathbf{l}$:

$$W_{ij}^{(2k+1)} = W_{ij}^{(2k)} \cdot \frac{a_i}{\sum_j W_{ij}^{(2k)}}, \qquad W_{ij}^{(2k+2)} = W_{ij}^{(2k+1)} \cdot \frac{l_j}{\sum_i W_{ij}^{(2k+1)}}. \tag{2.3}$$

Bacharach (1965) proves convergence to the unique matrix that minimizes the Kullback–Leibler divergence from the seed subject to the marginal constraints. The crucial implication is that the *topology* of the solution — which entries are zero and which are positive — is entirely determined by the seed matrix. IPF never creates new links; it only rescales existing ones. The quality of the reconstruction therefore depends critically on the quality of the seed.

This property is what makes IPF the natural second stage of our pipeline. The GNN provides an informed seed matrix that encodes learned structural patterns, and IPF then adjusts the weights to exactly satisfy the marginal constraints. The GNN determines *where* the money flows; IPF determines *how much*.

2.3 Graph Neural Networks

Graph Neural Networks (GNNs) extend deep learning to graph-structured data by learning node representations that aggregate information from local neighborhoods. We review the architectures relevant to our model, from the foundational message-passing framework to the attention-based mechanisms we employ.

2.3.1 Message Passing and Foundational Architectures

Gilmer et al. (2017) formalize the message-passing framework that unifies most GNN architectures. At each layer, a node i updates its representation by aggregating messages from its neighbors:

$$\mathbf{h}_i^{(\ell+1)} = \phi\left(\mathbf{h}_i^{(\ell)}, \bigoplus_{j \in \mathcal{N}(i)} \psi\left(\mathbf{h}_i^{(\ell)}, \mathbf{h}_j^{(\ell)}\right)\right), \tag{2.4}$$

where ψ is a message function, $\bigoplus$ is a permutation-invariant aggregation operator (sum, mean, or max), and ϕ is an update function. Different choices yield different architectures.

Kipf and Welling (2017) introduce Graph Convolutional Networks (GCN), which use a simplified spectral convolution with symmetric normalization. While foundational, GCN assigns equal weight to all neighbors, limiting its ability to capture the heterogeneous importance of different counterparties in financial networks. Hamilton et al. (2017) propose GraphSAGE, which introduces sampling-based aggregation and supports inductive learning on unseen nodes, a property relevant to our out-of-sample application where the 2024 bank population differs from the 2000–2005 training set. However, GraphSAGE still uses uniform or mean aggregation without learned attention.

Veličković et al. (2018) address this limitation with Graph Attention Networks (GAT), which learn attention coefficients that weight each neighbor's contribution differently:

$$\alpha_{ij} = \frac{\exp\big(\text{LeakyReLU}\big(\mathbf{a}^T[\mathbf{W}\mathbf{h}_i \| \mathbf{W}\mathbf{h}_j]\big)\big)}{\sum_{k \in \mathcal{N}(i)} \exp(\text{LeakyReLU}(\mathbf{a}^T[\mathbf{W}\mathbf{h}_i \| \mathbf{W}\mathbf{h}_k]))}. \tag{2.5}$$

Multi-head attention, where K independent attention heads are concatenated, allows the model to attend to different aspects of neighborhood structure simultaneously.

2.3.2 TransformerConv and the UniMP Architecture

Shi et al. (2021) propose the Unified Message Passing (UniMP) framework, which adapts the Transformer architecture of Vaswani et al. (2017) to graphs. Their TransformerConv layer computes multi-head attention using query, key, and value projections of node features:

$$\alpha_{ij}^{(h)} = \frac{(\mathbf{W}_Q^{(h)}\mathbf{h}_i)^T(\mathbf{W}_K^{(h)}\mathbf{h}_j)}{\sqrt{d_k}}, \qquad \mathbf{m}_{ij}^{(h)} = \alpha_{ij}^{(h)} \cdot \mathbf{W}_V^{(h)}\mathbf{h}_j, \tag{2.6}$$

where the superscript (h) indexes attention heads. Compared to GAT, TransformerConv provides several advantages for our application. The scaled dot-product attention is more expressive than the additive attention of GAT, the multi-head mechanism naturally supports learning different structural patterns in parallel, and the architecture includes an optional edge-feature channel that could incorporate bilateral attributes in future extensions.

We adopt a two-layer TransformerConv encoder with four attention heads in the first layer (producing $4 \times 64 = 256$-dimensional representations) and a single head in the second layer (projecting to 64 dimensions). This design allows the first layer to learn diverse neighborhood patterns while the second layer integrates them into a unified representation suitable for link prediction.

2.3.3 Link Prediction with GNNs

Link prediction — the task of estimating the probability that an edge exists between two nodes — is one of the canonical applications of GNNs (Zhang and Chen, 2018). The standard approach encodes each node into a latent vector $\mathbf{z}_i$ using a GNN encoder, then scores potential edges using a decoder that operates on pairs of node embeddings. Common decoders include the inner product $\sigma(\mathbf{z}_i^T\mathbf{z}_j)$, bilinear forms $\sigma(\mathbf{z}_i^T\mathbf{M}\mathbf{z}_j)$, and MLP-based decoders that process the concatenation or element-wise product of embeddings.

Kipf and Welling (2016) introduce the Variational Graph Auto-Encoder (VGAE), which frames link prediction as generative modeling: a GCN encoder maps nodes to a latent distribution, and the decoder reconstructs the adjacency matrix from sampled latent vectors. While

elegant, the VGAE's inner-product decoder is symmetric by construction, making it unsuitable for directed networks where the existence of a link from i to j does not imply a link from j to i.

Our decoder uses a dual-head MLP architecture. The link probability head takes the concatenation of source and target embeddings $[\mathbf{z}_i \| \mathbf{z}_j] \in \mathbb{R}^{128}$ and maps it through three linear layers ($128 \to 64 \to 32 \to 1$) with LayerNorm, LeakyReLU activations, and dropout. A separate reciprocity head takes the concatenation of both directional embeddings $[\mathbf{z}_i \| \mathbf{z}_j \| \mathbf{z}_j \| \mathbf{z}_i] \in \mathbb{R}^{256}$ and maps it through two linear layers ($256 \to 64 \to 1$), producing a residual correction that boosts the probability of edges likely to be reciprocated. The final logit is the sum of both outputs: $\ell_{ij} = \text{MLP}_{\text{prob}}([\mathbf{z}_i \| \mathbf{z}_j]) + \text{MLP}_{\text{recip}}([\mathbf{z}_i \| \mathbf{z}_j \| \mathbf{z}_j \| \mathbf{z}_i])$, with $p_{ij} = \sigma(\ell_{ij})$. This asymmetric design allows the model to assign different probabilities to the directed edges (i, j) and (j, i), which is essential for interbank networks where reciprocity is a meaningful structural property (approximately 62% of edges in our ground truth are reciprocated) rather than an automatic consequence of symmetry.

2.3.4 GNNs for Financial Network Reconstruction

The application of GNNs to interbank network reconstruction is recent and sparse. Petropoulos et al. (2020) apply machine learning techniques, including random forests and gradient-boosted trees, to reconstruct the Greek interbank network from partial data. While not using GNNs specifically, their work demonstrates that ML methods can outperform classical approaches on link prediction when historical data is available for training. However, their approach treats each potential edge independently, ignoring the graph structure that GNNs can exploit.

Our work differs from Petropoulos et al. in three respects. First, we use a GNN that operates on the graph structure itself, allowing information to propagate between nodes during encoding. Second, we integrate the GNN with IPF to guarantee marginal consistency, which their approach does not. Third, we train on structural losses (reciprocity, degree distribution, disassortativity) in addition to edge classification, explicitly targeting the microstructural properties that matter for contagion analysis.

2.4 The e-MID Platform

The electronic Market for Interbank Deposits (e-MID) is a screen-based trading platform for unsecured interbank deposits, operated by e-MID S.p.A. and based in Milan. Active since 1990, e-MID was the only electronic platform for unsecured interbank lending in the euro area for most of its history, making it a unique source of transaction-level data on interbank relationships. The platform records every transaction with timestamp, lender identity, borrower identity, amount, maturity, and interest rate.

During our training period (2000–2005), e-MID captured approximately 60–70% of the Italian unsecured overnight interbank market, with the remainder conducted over the counter or through bilateral agreements. The platform's significance diminished after the 2007–2008 financial crisis as banks shifted toward secured (repo) lending and ECB facilities, but the pre-crisis data remains the most complete publicly available record of an interbank network with bilateral resolution.

The e-MID data has been extensively studied. Precup et al. (2005) and Iori et al. (2008) establish the empirical regularities of the Italian overnight market that inform our model design: heavy-tailed degree distributions, disassortative mixing, moderate clustering, and a stable core–periphery structure. Fricke and Lux (2015) formalize the core–periphery analysis using

stochastic block models and document that approximately 25–30 banks form a persistent core that intermediates between a periphery of 100–150 less active institutions. These structural features are the targets our GNN is designed to learn and reproduce.

2.5 Contagion Modeling: From Clearing to DebtRank

2.5.1 The Eisenberg–Noe Clearing Framework

Eisenberg and Noe (2001) model the interbank system as a network of obligations. Bank i owes a total of $\bar{p}_i$ to its creditors, distributed according to a relative liability matrix Π where Π_{ij} is the fraction of i's obligations owed to j. Each bank pays out of its available resources: operating assets e_i plus payments received from other banks. A clearing vector $\mathbf{p}^*$ satisfies

$$p_i^* = \min\left(\bar{p}_i,\ e_i + \sum_j \Pi_{ji}\, p_j^*\right), \tag{2.7}$$

ensuring that solvent banks pay in full while insolvent banks pay pro rata. The framework provides a static equilibrium analysis of default cascades: given a shock to bank assets, one computes the new clearing vector and counts which banks default.

While foundational, the Eisenberg–Noe framework has a well-known limitation: it models only direct contagion through contractual obligations, ignoring amplification through mark-to-market losses, fire sales, funding withdrawals, or confidence effects. This underestimates systemic risk, particularly in networks where banks hold each other's equity or debt instruments.

2.5.2 DebtRank

Battiston et al. (2012) introduce DebtRank to address this limitation. Unlike the Eisenberg–Noe clearing mechanism, which propagates losses only upon actual default, DebtRank models *distress propagation*: a bank experiencing partial equity losses transmits proportional distress to its creditors even if it remains solvent. The original formulation tracks the relative equity loss $h_i(t) \in [0,1]$ of each bank across discrete time steps, with three states: *undistressed* ($h_i = 0$), *distressed* ($0 < h_i < 1$), and *defaulted* ($h_i = 1$). At each step, distress propagates from newly distressed banks to their creditors through the leverage matrix $\Lambda_{ij} = W_{ij}/E_i$, where W_{ij} is i's exposure to j and E_i is i's equity. A key feature of the original DebtRank is that each bank can propagate distress only *once*: after transmitting losses, it becomes *inactive* and ceases to be a source of further contagion. This one-shot mechanism prevents cycles and ensures convergence in at most n steps, but it also truncates the cascading dynamics.

Bardoscia et al. (2015) provide a "microscopic foundation" for DebtRank by removing the one-shot constraint and deriving the dynamics from first principles. In their differential formulation, distress propagates continuously:

$$h_i(t+1) = \min\left(1,\ h_i(t) + \sum_j \Lambda_{ij}\, \Delta h_j(t)\right), \tag{2.8}$$

where $\Delta h_j(t) = h_j(t) - h_j(t-1)$ is the *incremental* distress at bank j. Only the increment propagates at each step, avoiding double-counting while allowing multi-round amplification. This formulation converges to a fixed point where no further distress increments occur, producing

generally higher systemic risk estimates than the original one-shot version. The differential DebtRank is the formulation adopted by Gurgone and Iori (2022) and the one we implement in this thesis.

The aggregate DebtRank of the system is computed by running the dynamics for each bank z as the initial defaulter ($h_z(0) = 1$, $h_i(0) = 0$ for $i \neq z$) and measuring the resulting system-wide equity loss:

$$R_z = \frac{\sum_{i \neq z} h_i(T) \cdot E_i}{\sum_j E_j}, \tag{2.9}$$

where T is the convergence time. The value R_z measures the fraction of total system equity destroyed by the default of bank z alone, providing a bank-level measure of systemic importance that incorporates network position, counterparty structure, and cascade dynamics.

2.6 Macroprudential Capital Buffers

The Basel III framework, implemented in the European Union through the Capital Requirements Directive IV (CRD IV), introduces macroprudential capital buffers that require systemically important institutions to hold additional Common Equity Tier 1 (CET1) capital above the minimum Pillar 1 requirement of 8% of risk-weighted assets. The rationale is that institutions whose failure would impose disproportionate costs on the financial system should internalize this externality through higher capital requirements.

The identification of Other Systemically Important Institutions (O-SIIs) at the national level follows the guidelines of European Banking Authority (2014), which prescribe a scoring methodology based on four equally weighted criteria: size (total assets), importance (payment activity and private-sector deposits), complexity and cross-border activity, and interconnectedness (interbank assets, liabilities, and debt securities). Each bank receives a composite score between 0 and 100, and national authorities assign buffer surcharges (typically 0–2% of RWA) based on score thresholds.

Gurgone and Iori (2022) provide the methodological framework for our policy application. Using an agent-based model (ABM) of a heterogeneous banking network with endogenous interbank lending and liquidity crises, they compare EBA indicator-based buffer allocation with alternative allocations based on DebtRank scores. Their key finding is that DebtRank-based scoring, which accounts for network position and contagion dynamics, can achieve the same reduction in systemic risk with less total additional capital than the standard EBA approach, because it targets buffers more precisely at the institutions that contribute most to systemic fragility. They also find that the correlation between EBA scores and DebtRank impact scores is low, suggesting that size-based indicators are imperfect proxies for systemic importance.

Our approach differs from Gurgone and Iori in an important respect. Their analysis uses an ABM with endogenous network formation and dynamic balance sheet adjustment, while ours applies the scoring framework to a statically reconstructed network. This means our results capture the *structural* effect of buffer allocation — how the topology of exposures shapes systemic risk under different capital regimes — but not the *behavioral* feedback effects that would arise if banks adjusted their lending in response to changed capital requirements. We implement the deleveraging channel — undercapitalized banks reduce interbank lending proportionally to their capital shortfall — but do not model the second-round effects of this deleveraging on other banks' funding or the endogenous rewiring of lending relationships. This limitation is inherent to the static reconstruction approach and is discussed further in Chapter 7.

Chapter 3

Data

This chapter describes the three data sources used in this thesis: the e-MID transaction-level dataset that provides training and validation data for the GNN, the macroeconomic variables from European Central Bank statistical services that augment the node feature vectors, and the ORBIS balance sheet database that supplies the cross-sectional information required for the 2024 policy application. We detail the cleaning procedures, the construction of daily graph objects, the feature engineering strategy, and the bank classification scheme used throughout the analysis.

3.1 The e-MID Transaction Dataset

Our primary data source is the complete record of overnight interbank transactions on the e-MID platform from January 2000 to December 2009. Each transaction record contains the date, the anonymized identifiers of the lending and borrowing institutions, the transaction amount (in millions of euros), the interest rate, and the maturity. We restrict attention to Italian banks, filtering on the country prefix of the institution codes, and to overnight maturities, which constitute the dominant segment of the platform and the focus of the microstructural literature (Precup et al., 2005; Iori et al., 2008).

After filtering, the dataset comprises approximately 1.1 million individual transactions involving 218 distinct Italian banks over the ten-year period. The number of active banks varies over time, with some institutions entering or exiting the platform across years, but the set of 218 institutions represents the union of all banks that appear at least once during the sample period. We maintain a fixed node set throughout, assigning zero activity to banks in periods where they are inactive. This design choice ensures dimensional consistency across daily graph objects and avoids the complications of dynamic node sets in the GNN architecture.

3.1.1 Temporal Structure

We partition the data into three segments based on their role in the pipeline:

- **Training period (2000–2005):** 1,565 daily transaction graphs used to train the GNN. This period covers a relatively stable phase of the Italian interbank market, preceding the onset of the global financial crisis.
- **Test period (2006–2007):** 521 daily graphs used for out-of-sample evaluation. The bi-annual aggregate of this period serves as the ground truth matrix against which reconstruction methods are compared.

- **Excluded period (2008–2009):** The crisis years are excluded from both training and testing. Including them in training would require a post-2009 test period that is unavailable, and inserting crisis data between a pre-crisis training set and a pre-crisis test set would create temporal leakage. The e-MID platform also experienced a significant contraction in activity during this period as banks shifted toward secured lending and ECB facilities, making the data less representative of normal interbank market functioning.

3.2 Daily Graph Construction

For each business day t in the sample, we construct a directed weighted graph $G_t = (V, E_t, \mathbf{X}_t)$ where V is the fixed set of 218 nodes, E_t is the set of directed edges observed on day t, and $\mathbf{X}_t \in \mathbb{R}^{218\times 17}$ is the node feature matrix.

An edge $(i, j) \in E_t$ exists if bank i lent to bank j at least once on day t. The edge weight $w_{ij}^{(t)}$ is the total amount lent from i to j on that day, aggregated across all transactions between the pair. Self-loops are excluded ($w_{ii}^{(t)} = 0$). The resulting daily graphs are sparse: a typical day has between 300 and 600 directed edges among 218 nodes, corresponding to a density of approximately 0.6–1.3%.

The edge index of each graph object is stored as the ground truth label for link prediction training. Crucially, these edges are *never* provided to the GNN encoder as input; the encoder sees only the node feature matrix and a gravity-based scaffold constructed from lagged features. This separation between input structure and prediction target is essential to prevent information leakage, as discussed in Section 3.3.

3.3 Node Feature Engineering

Each node i on day t is described by a 17-dimensional feature vector $\mathbf{x}_i^{(t)} \in \mathbb{R}^{17}$. The features are organized into three groups, each designed to provide the GNN with informative signals about a bank's role in the interbank network without leaking information about the edges it is tasked with predicting.

3.3.1 Lagged Annual Activity Features (4 dimensions)

The first four features capture each bank's historical activity level on the e-MID platform, all derived from the preceding calendar year $Y-1$:

1. $\log(1+V_i^{\text{lent},Y-1})$: log-transformed total amount lent by bank i during the preceding calendar year $Y-1$.
2. $\log(1+V_i^{\text{borrowed},Y-1})$: log-transformed total amount borrowed by bank i during $Y-1$.
3. $\log(1+V_i^{\text{total},Y-1})$: log-transformed total interbank volume (lending plus borrowing) during $Y-1$.
4. r_i^{Y-1}: normalized size rank based on $V_i^{\text{total},Y-1}$, computed as $1-(\text{rank}_i-1)/\max(\text{rank})$ so that the most active bank receives a value near 1 and the least active near 0.

The one-year lag ensures strict temporal separation: when predicting edges on any day in year Y, the model uses volume information from $Y-1$ only. No same-year transaction data enters the features. The log transformation compresses the heavy-tailed distribution of interbank volumes and stabilizes the scale across banks of different sizes. Feature [3] provides an ordinal summary of each bank's position in the activity hierarchy that is invariant to the absolute scale of transaction volumes, complementing the cardinal measures in features [0]–[2].

Features [0] and [1] serve a dual purpose. In addition to informing the GNN about each bank's activity level, they are used to construct the gravity scaffold that provides the encoder with its input graph structure (Section 4.1). This design creates a clean information architecture: the scaffold is a deterministic function of lagged volumes, and the encoder's task is to refine the scaffold's coarse connectivity pattern into a precise edge probability map.

3.3.2 Institutional Group Dummies (5 dimensions)

Features [4] through [8] are one-hot indicators for the bank's size classification according to Banca d'Italia thresholds based on total assets:

Group	Total assets threshold	Banks in sample
MA (Maggiori)	>60 billion EUR	5
GR (Grandi)	26–60 billion EUR	7
ME (Medi)	9–26 billion EUR	16
MI (Minori)	1.3–9 billion EUR	65
PI (Piccoli)	<1.3 billion EUR	125

These group indicators are time-invariant within our sample and capture the persistent structural role of each bank in the interbank network. The core–periphery structure documented by Fricke and Lux (2015) largely aligns with this size classification: MA and GR banks constitute the network core, while MI and PI banks form the periphery.

3.3.3 Macroeconomic Contextual Variables (8 dimensions)

Features [9] through [16] are market-wide macroeconomic and monetary indicators that capture the evolving conditions under which interbank trading occurs. These variables are common to all nodes on a given day but vary across time, providing the GNN with a temporal context signal:

9. ECB Main Refinancing Operations (MRO) rate
10. M3 money supply growth (annual, %)
11. EURIBOR 3-month rate
12. HICP inflation (Harmonised Index of Consumer Prices, %)
13. Italian unemployment rate (monthly, forward-filled daily)
14. BTP–Bund spread (Italian 10-year yield minus German 10-year yield, measuring sovereign risk)

15. Quarterly GDP growth (%, forward-filled daily)

16. Recession indicator (binary: 1 if quarterly GDP growth is negative, 0 otherwise)

These variables are sourced from the ECB Statistical Data Warehouse and the OECD Main Economic Indicators database. Rates and yields are recorded at daily frequency where available; the unemployment rate, available monthly, is forward-filled to daily frequency. The recession indicator is derived mechanically from the GDP growth series. All variables are contemporaneous with the prediction date — they do not require lagging because they are publicly available macroeconomic indicators, not bank-specific information derived from the e-MID data.

The rationale for including macro variables is twofold. First, the structure of interbank trading varies systematically with monetary policy conditions: when the MRO rate rises, the volume and pattern of overnight lending adjusts as banks' liquidity needs change. Second, the EONIA–MRO spread is a well-established indicator of interbank market stress (Iori et al., 2008), and its inclusion allows the GNN to learn that network structure differs between calm and stressed periods, even within the pre-crisis training sample.

3.4 Leakage Prevention

The feature engineering strategy is designed to prevent three forms of information leakage:

Direct leakage. No same-day e-MID transaction data appears in the feature vector. The volume features use data from prior calendar years only. The macro variables are public information, not derived from e-MID.

Scaffold leakage. The gravity scaffold (Section 4.1) is constructed from features [0] and [1], which are lagged annual volumes. While the scaffold edges will overlap with true edges (approximately 90% overlap in our verification tests, because banks that were active last year tend to be active this year), this overlap reflects genuine predictive signal from historical activity patterns, not unauthorized access to contemporaneous edge labels.

Temporal leakage. The train/test split is strictly temporal: the model trains on 2000–2005 and is evaluated on 2006–2007. No future information enters the training process. Within the training set, the 85/15 train/validation split is also temporal (the last 15% of training days serve as validation), ensuring that the early stopping criterion is evaluated on genuinely held-out future data.

We verify the absence of leakage empirically. For a sample test graph, the scaffold contains 5,750 edges while the true graph contains 326 edges. The overlap of 295 edges (90.5% of true edges) reflects the persistence of interbank relationships across years, not information contamination. The remaining 5,455 scaffold edges that do not correspond to true edges are the candidates that the GNN must learn to distinguish from genuine links.

3.5 ORBIS Balance Sheet Data for 2024

The counterfactual policy analysis in Chapter 6 requires balance sheet data for the current Italian banking system. We obtain this from Bureau van Dijk's ORBIS Bank Focus database, extracting the most recent available annual filings (fiscal year 2023 or 2024) for all Italian commercial banks, cooperative banks, and savings banks.

3.5.1 Sample Construction

The initial extraction yields 343 Italian banking institutions. We apply the following filters:

1. **Exclusion of non-commercial entities.** We remove Banca d'Italia (the central bank), Cassa Depositi e Prestiti (the state development bank), and Cassa di Compensazione e Garanzia (the central counterparty), as these institutions do not participate in the commercial interbank market in a manner comparable to ordinary banks. After this filter, 340 banks remain.

2. **Data completeness.** We require non-missing values for total assets, total equity, and either interbank assets or total loans. Banks with missing equity or zero total assets are dropped.

3. **Activity threshold.** Banks with total assets below 100 million EUR are excluded as too small to participate meaningfully in the interbank market.

The final sample comprises 203 Italian banks. For each bank, we extract or compute the following variables:

Variable	Description
Total assets	From ORBIS, in millions EUR
Total equity	Shareholders' equity from ORBIS
CET1 ratio	Where reported; otherwise estimated as equity / (total assets $\times$ 0.5)
Interbank assets	Loans and advances to banks (proxy for total interbank lending)
Interbank liabilities	Due to banks (proxy for total interbank borrowing)
RWA (approx.)	Estimated as total assets $\times$ 0.5, following the average ratio observed in Italian banks' Pillar 3 disclosures

The RWA approximation of 50% of total assets is a simplification. The true risk-weight density varies across banks depending on asset composition, with retail-heavy banks typically having lower ratios and investment-heavy banks having higher ones. However, individual RWA data is not systematically available in ORBIS for all banks in our sample. The 50% ratio is consistent with the aggregate risk-weight density reported by the EBA for Italian banks in its 2023 transparency exercise and provides a reasonable approximation for the purpose of computing capital adequacy ratios.

3.5.2 Bank Classification

We classify banks into five size groups following the Banca d'Italia's institutional categorization based on total assets (Banca d'Italia, 2023):

Group	Threshold	Banks	Mean assets (bn EUR)	Mean CET1
MA (Maggiori)	>60 bn	15	206.4	15.8%
GR (Grandi)	26–60 bn	5	37.2	15.5%
ME (Medi)	9–26 bn	22	14.8	18.6%
MI (Minori)	1.3–9 bn	100	3.2	20.3%
PI (Piccoli)	<1.3 bn	198	0.6	29.7%

The classification above applies to all 340 banks in the ORBIS extraction. After applying the interbank-activity threshold (banks with non-negligible interbank assets or liabilities), 203 banks remain and form the nodes of the reconstructed network. The excluded institutions are predominantly small (PI) banks with no reported interbank lending or borrowing. The temporal policy analysis in Chapter 6, which requires balance sheet data from both 2023 and 2024, further restricts the sample to the 188 banks present in both years' filings. The 15 banks lost to this intersection requirement are predominantly small institutions that entered or exited the ORBIS database between filings.

The inverse relationship between bank size and CET1 ratio is consistent with the broader European pattern: smaller banks tend to maintain higher capital ratios, partly because they face higher idiosyncratic risk and partly because they have less access to wholesale funding markets and therefore hold more equity as a buffer. The 15 MA banks in the 2024 sample account for over 75% of total system assets, reflecting the high concentration of the Italian banking sector following the consolidation wave of the 2010s.

Note that the number of banks in each group differs between the e-MID sample (218 banks, 2000–2005) and the ORBIS sample (203 banks, 2024). This reflects both genuine changes in the Italian banking landscape — mergers, acquisitions, and the disappearance of small cooperative banks — and differences in data coverage between the two sources. The GNN handles this discrepancy through the feature bridge described in Section 3.6.

3.6 Feature Bridge: e-MID to ORBIS

Applying the GNN trained on e-MID data to the 2024 ORBIS sample requires mapping the 17-dimensional feature vector from the training domain to the inference domain. The mapping is constructed as follows:

Dim	Training (e-MID)	Inference (ORBIS)	Rationale
0	$\log(1+\text{annual lent}_{Y-1})$	$\log(1+\text{IB assets})$	Flow $\leftrightarrow$ stock proxy
1	$\log(1+\text{annual borrowed}_{Y-1})$	$\log(1+\text{IB liabilities})$	Flow $\leftrightarrow$ stock proxy
2	$\log(1+\text{total volume}_{Y-1})$	$\log(1+\text{IB assets}+\text{IB liabilities})$	Sum of [0] and [1] in
3	Normalized size rank by $Y-1$ volume	Normalized size rank by total assets	Ordinal activity proxy
4–8	Group dummies	Group dummies	Same classification sc
9–15	Macro variables	Set to 2024 averages	Public data
16	Recession indicator	From 2024 GDP data	Derived from [15]

The most consequential mapping is features [0]–[1]: annual e-MID transaction volumes (a flow measure) are replaced by ORBIS interbank assets and liabilities (a stock measure). These quantities are conceptually analogous — both capture the scale of a bank's interbank activity — but numerically different in distribution. The log transformation partially mitigates

the distributional shift, but we acknowledge that the GNN's learned representations may not transfer perfectly across this domain gap. The gravity scaffold, which uses features [0]–[1] to construct the input graph, will similarly reflect balance sheet stocks rather than transaction flows, potentially producing a denser or sparser scaffold than the GNN encountered during training.

Features [2] and [3] have natural ORBIS analogues. Feature [2], log total interbank volume, is computed as $\log(1 + \text{IB assets} + \text{IB liabilities})$, directly paralleling the training-time construction $\log(1 + V_i^{\text{lent}} + V_i^{\text{borrowed}})$. Feature [3], the normalized size rank, is computed by ranking banks by total assets rather than by interbank volume; this substitution is necessary because interbank volume alone is a noisy proxy for relative size in the ORBIS cross-section, whereas total assets provides a stable ordering consistent with the Banca d'Italia classification. The macro variables [9]–[16] are set to their average values over 2024 using the same ECB sources as the training data.

This feature bridge is an inherent limitation of applying a model trained on historical microdata to a contemporary cross-section. The alternative — retraining the GNN on recent transaction data — would require access to current bilateral transaction records, which is precisely the information the reconstruction pipeline is designed to circumvent. The bridge therefore represents a pragmatic compromise that leverages the structural patterns learned from historical data while adapting to the current institutional landscape through the available balance sheet proxies.

3.7 Feature Bridge Sensitivity

The most consequential element of the feature bridge is the flow-to-stock substitution in features [0]–[1], which maps annual e-MID transaction volumes to ORBIS balance sheet positions. Features [2] and [3] introduce a secondary concern: feature [2] is mechanically derived from [0] and [1] in both domains (it is their log-sum), so any distortion in [0]–[1] propagates; feature [3] substitutes an asset-based rank for a volume-based rank, introducing a conceptual rather than merely distributional shift.

To assess the sensitivity of model predictions to features [2]–[3], we conduct an ablation in which these two features are zeroed out at inference time while the remaining 15 features are held at their original values.

Zeroing features [2]–[3] shifts the predicted edge probabilities substantially: the mean probability across all candidate pairs drops from 0.128 to 0.027 and the standard deviation contracts from 0.234 to 0.059. The probability mass shifts downward because the model interprets the absence of total volume and rank information as a signal of low activity, reducing estimated link probabilities across the board.

Despite this magnitude shift, the *ranking* of edge probabilities is largely preserved. The Spearman rank correlation between the original and zeroed probability vectors is $\rho = 0.947$, and the Pearson correlation between the raw vectors is 0.670. Because the pipeline uses edge probabilities only for top-k selection — a rank-based operation — the zeroing of features [2]–[3] has a limited effect on the reconstructed network topology. The same edges would be selected under either feature specification, since their relative ordering is nearly identical.

However, the probability magnitudes themselves are not robust to this perturbation. Any application that uses the GNN's predicted probabilities as calibrated confidence scores — rather than as an ordinal ranking for edge selection — would be sensitive to the feature bridge design. The edge-budget sensitivity analysis in Chapter 7 further examines the downstream con-

sequences by varying the selection threshold, and the alternative feature mapping robustness check (Chapter 7, Section 7.6.2) replaces the absolute-volume bridge with a relative-exposure specification ($\log(1 + \text{IB assets}/\text{total assets})$), confirming that the qualitative conclusions of the policy analysis are preserved.

Chapter 4

Methodology

This chapter presents the complete methodological pipeline in six parts: the gravity scaffold that provides the GNN with an input topology, the GNN architecture and its multi-objective training procedure, the IPF integration that enforces marginal consistency, the three baseline reconstruction methods used for comparison, the differential DebtRank algorithm used for systemic risk assessment, and the capital buffer scoring and scenario construction framework used for the policy application.

Figure 4.1 summarizes the pipeline schematically. Given a set of banks with observed marginals (total interbank lending and borrowing), the gravity scaffold produces a dense candidate graph. The GNN encoder processes this graph and outputs edge probabilities for all node pairs. The top-k most probable edges are selected, weighted by a gravity kernel, and passed as a seed matrix to IPF, which iteratively rescales the weights to match the target marginals exactly. The output is a weighted directed adjacency matrix $\hat{W}$ that can be used for contagion analysis.

4.1 Gravity Scaffold Construction

The GNN encoder requires an input graph to perform message passing. In a standard GNN application, this input graph is the observed network itself. In our setting, the observed network is precisely what we are trying to reconstruct, so providing it as input would constitute information leakage. We resolve this by constructing a *gravity scaffold*: a synthetic input graph derived entirely from lagged node features, independent of the true edge set.

The construction follows the gravity model tradition in trade and migration economics (Tinbergen, 1962), adapted to directed interbank networks. For each node i, let $s_i = \exp(x_{i,0}) - 1$ and $b_i = \exp(x_{i,1}) - 1$ be the inverse log-transformed values of features [0] and [1], recovering the original lagged annual lending and borrowing volumes. The gravity score for the directed pair (i, j) is

$$g_{ij} = s_i \cdot b_j, \tag{4.1}$$

measuring the product of i's lending propensity and j's borrowing propensity. This formulation is inherently disassortative by construction: a large lender paired with a large borrower receives a high score regardless of whether the two institutions are similar in overall size, reproducing the core–periphery connectivity pattern where major lenders intermediate funds to major borrowers while peripheral banks appear on one side or the other.

The scaffold connects each node to its top-k counterparties by gravity score, applied in both directions: each potential lender i is connected to its k highest-scoring borrowers, and each

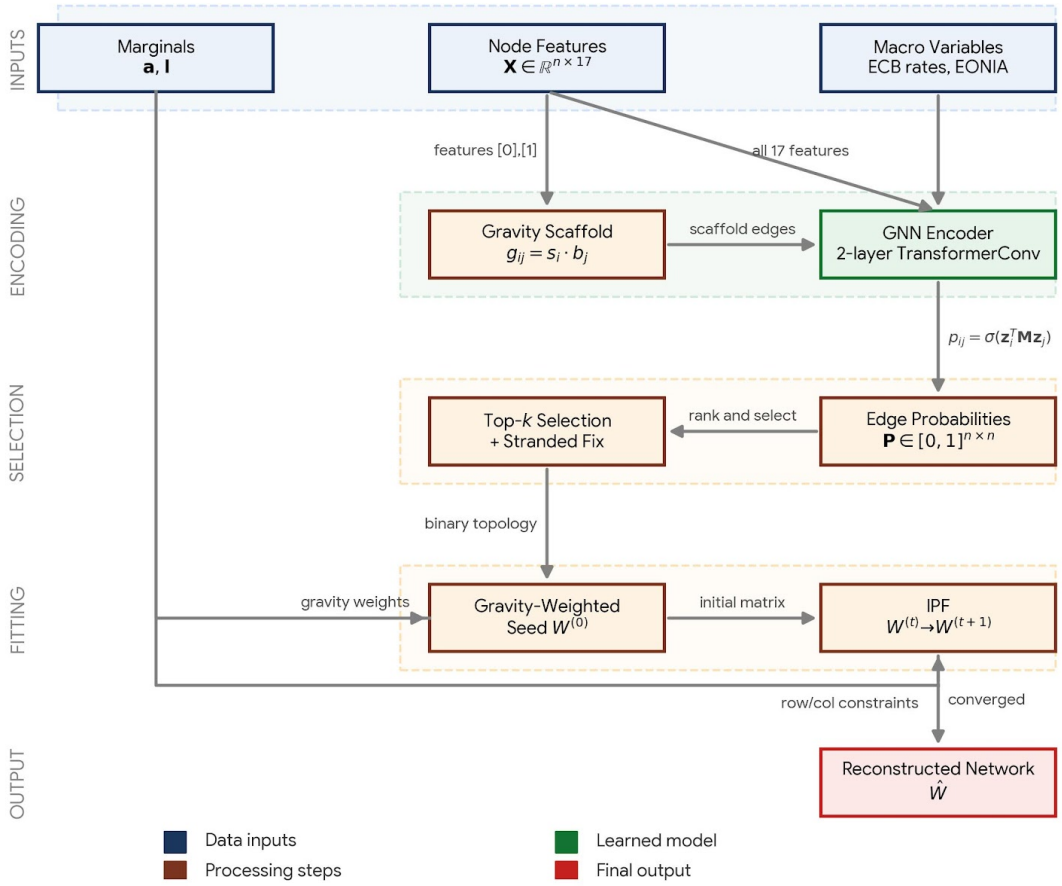

Figure 4.1: GNN+IPF pipeline overview. Node features and macroeconomic variables are processed by a gravity scaffold (using lagged volumes) and a two-layer TransformerConv encoder to produce edge probabilities. The top-k most probable edges are selected and weighted by a gravity kernel, then passed as a seed matrix to IPF, which iteratively rescales the weights to match the observed marginals $\mathbf{a}$ (total interbank assets) and $\mathbf{l}$ (total interbank liabilities). The GNN determines the network topology; IPF determines the edge weights.

potential borrower j is connected to its k highest-scoring lenders. Formally, the scaffold edge set is

$$E_{\text{scaffold}} = \{(i,j) : j \in \text{top-}k_j(g_{i,\cdot})\} \cup \{(i,j) : i \in \text{top-}k_i(g_{\cdot,j})\}. \tag{4.2}$$

Self-loops are excluded. The union of lender-side and borrower-side top-k sets produces a scaffold that is denser than either alone, ensuring that peripheral banks with low lending or borrowing volumes are still connected to the core.

The hyperparameter k controls the scaffold density. With $k = 30$ and $n = 218$ nodes, the scaffold contains approximately 5,700–5,800 directed edges, corresponding to a density of roughly 12%. This is substantially denser than the true daily graphs (density 0.6–1.3%) but deliberately so: the scaffold's purpose is to provide the GNN with a sufficiently rich neighborhood structure for message passing, not to approximate the true network. The GNN's task is to assign low probabilities to most scaffold edges and high probabilities to the subset that corresponds to genuine interbank links.

The scaffold is deterministic given the node features and requires no trainable parameters. At inference time, it is constructed identically from the ORBIS-derived features, ensuring methodological consistency between training and application.

4.2 GNN Architecture

The GNN consists of a TransformerConv encoder that maps node features to latent representations and a dual-head MLP decoder that scores potential edges.

4.2.1 Encoder

Layer 1. The raw features are passed through a LayerNorm operation, then processed by a TransformerConv layer with four attention heads and a per-head output dimension of 64, producing a $4 \times 64 = 256$-dimensional concatenated representation. This is followed by LayerNorm, an ELU activation, and dropout (rate 0.3).

Layer 2. The 256-dimensional representations are processed by a second TransformerConv layer with a single attention head and output dimension 64, followed by a final LayerNorm:

$$\mathbf{z}_i = \text{LayerNorm}\Big(\text{TransformerConv}^{(2)}\Big(\text{ELU}\Big(\text{LayerNorm}\Big(\text{TransformerConv}^{(1)}(\mathbf{x}_i, E_{\text{scaffold}})\Big)\Big), E_{\text{scaffold}}\Big)\Big) \in \mathbb{R} \tag{4.3}$$

Both TransformerConv layers use the learnable skip-connection parameter β from the UniMP architecture (Shi et al., 2021), which interpolates between the aggregated neighborhood message and the node's own transformed features.

4.2.2 Decoder

The decoder assigns a probability to each potential directed edge (i,j) by summing the outputs of two MLP heads. The **link probability head** takes the concatenation of source and target embeddings $[\mathbf{z}_i \| \mathbf{z}_j] \in \mathbb{R}^{128}$ and maps it through three linear layers ($128 \to 64 \to 32 \to 1$) with LayerNorm, LeakyReLU activations, and dropout. The **reciprocity head** takes the concatenation of both directional embeddings $[\mathbf{z}_i \| \mathbf{z}_j \| \mathbf{z}_j \| \mathbf{z}_i] \in \mathbb{R}^{256}$ and maps it through two linear layers ($256 \to 64 \to 1$) with LayerNorm, LeakyReLU, and dropout. The final logit is the sum of both outputs:

$$\ell_{ij} = \text{MLP}_{\text{prob}}([\mathbf{z}_i \| \mathbf{z}_j]) + \text{MLP}_{\text{recip}}([\mathbf{z}_i \| \mathbf{z}_j \| \mathbf{z}_j \| \mathbf{z}_i]), \qquad p_{ij} = \sigma(\ell_{ij}). \tag{4.4}$$

The reciprocity head acts as a residual correction: by processing both orderings (i, j) and (j, i) simultaneously, it can boost the probability of edges that are likely to be reciprocated. Both heads are active during training and inference. Weights are initialized using Kaiming normal initialization with LeakyReLU nonlinearity.

4.3 Multi-Objective Training

The model is trained to minimize a composite loss function that combines edge classification accuracy with three structural objectives targeting the microstructural properties identified by Precup et al. (2005). The total loss is

$$\mathscr{L} = \lambda_{\text{cls}}\mathscr{L}_{\text{cls}} + \lambda_{\text{recip}}\mathscr{L}_{\text{recip}} + \lambda_{\text{deg}}\mathscr{L}_{\text{deg}} + \lambda_{\text{disassort}}\mathscr{L}_{\text{disassort}}, \tag{4.5}$$

with weights $\lambda_{\text{cls}} = 1.0$, $\lambda_{\text{recip}} = 0.02$, $\lambda_{\text{deg}} = 0.5$, and $\lambda_{\text{disassort}} = 0.5$.

4.3.1 Edge Classification Loss

The primary loss is binary cross-entropy over sampled positive and negative edges:

$$\mathscr{L}_{\text{cls}} = -\frac{1}{|S|} \sum_{(i,j)\in S} \left[y_{ij} \log p_{ij} + (1 - y_{ij}) \log(1 - p_{ij}) \right], \tag{4.6}$$

where S is a balanced sample of positive edges (from the true edge set) and negative edges (randomly sampled non-edges), with a 1:1 positive-to-negative ratio. The balanced sampling addresses the class imbalance inherent in sparse graphs, where non-edges vastly outnumber edges.

4.3.2 Reciprocity Loss

The reciprocity loss encourages the model to assign high probabilities to edges whose reverse also exists in the training graph. For each sampled pair (i, j) that is reciprocated in the ground truth — that is, where both $(i, j) \in E_t$ and $(j, i) \in E_t$ — we apply binary cross-entropy with a target of 1:

$$\mathscr{L}_{\text{recip}} = -\frac{1}{|R|} \sum_{(i,j)\in R} \log \sigma(\ell_{ij}), \tag{4.7}$$

where $R = \{(i, j) \in S : (j, i) \in E_t\}$ is the subset of sampled pairs whose reverse exists in the true edge set and ℓ_{ij} is the combined logit from Equation (4.4). This loss operates on the joint output of both decoder heads, meaning it trains the reciprocity head and the link probability head jointly: the reciprocity head learns to boost the logit for reciprocated pairs, while the link probability head learns not to suppress them.

4.3.3 Degree Distribution Loss

The degree distribution loss measures the discrepancy between the predicted and true degree distributions. For each training graph, we compute the log-degree of each node from the predicted probabilities and compare it to the log-degree from the true edges:

$$\mathscr{L}_{\text{deg}} = \frac{1}{n} \sum_{i=1}^{n} \left(\log(1 + \hat{d}_i) - \log(1 + d_i) \right)^2, \tag{4.8}$$

where $d_i = \sum_j \mathbf{1}[(i,j) \in E_t]$ is the true out-degree and $\hat{d}_i = \sum_j p_{ij}$ is the expected out-degree under the predicted probabilities. The log transformation ensures that the loss is not dominated by the high-degree core nodes and gives appropriate weight to the degree distribution's tail, which is critical for reproducing the power-law-like shape documented by Iori et al. (2008).

Note that $\hat{d}_i$ is computed as $\sum_{j:(i,j)\in S} p_{ij}$, where S is the set of sampled pairs, not the full set of $n(n-1)$ potential edges. This stochastic approximation is computationally necessary (evaluating all pairs would require $O(n^2)$ forward passes per training step) but introduces noise into the degree estimate, particularly for high-degree nodes whose sampled neighborhood is a small fraction of their true connectivity.

4.3.4 Disassortativity Loss

The disassortativity loss penalizes the model for assigning high probabilities to edges between nodes of similar degree. Node degrees are computed from the ground truth edge set. For each sampled pair $(i,j) \in S$ (both positive and negative), the penalty is

$$\mathscr{L}_{\text{disassort}} = \frac{1}{|S|} \sum_{(i,j)\in S} p_{ij} \cdot \left(1 - \frac{|d_i - d_j|}{d_{\max}}\right), \tag{4.9}$$

where d_i and d_j are the true total degrees of the source and target nodes and $d_{\max}$ is the maximum degree in the graph. The term $(1 - |d_i - d_j|/d_{\max})$ is close to 1 when the two nodes have similar degree (assortative pair) and close to 0 when they have very different degree (disassortative pair). Multiplying by p_{ij} means that the penalty is incurred only when the model assigns high probability to an assortative edge.

Note that this loss operates over all sampled pairs, not only positive edges. The degree reference d_i is computed from the ground truth edges of the current training graph, creating a dependency on the true labels within the loss computation. This is a form of teacher forcing: the model is guided toward the degree structure of the real network during training. A potential inconsistency is that the degree distribution loss simultaneously adjusts the predicted degree, creating a moving-target dynamic that we do not attempt to resolve.

4.3.5 Optimization

The model is optimized using AdamW (Loshchilov and Hutter, 2019) with an initial learning rate of 10^{-3} and weight decay of 10^{-5}. A ReduceLROnPlateau scheduler reduces the learning rate by a factor of 0.5 if the validation loss does not improve for 3 consecutive epochs. Early stopping with a patience of 30 epochs terminates training when the validation loss ceases to decrease, using the last 15% of training days (235 graphs) as the validation set. Gradients are clipped to a maximum ℓ_2 norm of 1.0 to prevent exploding gradients.

In practice, the model converges rapidly: training terminates at epoch 8 with a test AUROC of 0.9255 and test loss of 0.4202. The fast convergence reflects the strong signal in the gravity scaffold combined with the relatively simple structural patterns in the interbank data.

4.4 From GNN Probabilities to Weighted Networks

The GNN produces a matrix of edge probabilities $\mathbf{P} \in [0,1]^{n\times n}$ for all directed node pairs. To obtain a weighted adjacency matrix $\hat{W}$ consistent with the observed marginals $\mathbf{a}$ (row sums) and $\mathbf{l}$ (column sums), we apply a three-step procedure.

4.4.1 Edge Selection

Given a target number of edges m (set equal to the observed number of edges in the ground truth period, or estimated from historical density for out-of-sample application), we select the m pairs with the highest predicted probabilities:

$$\hat{E} = \text{top-}m\{(i,j) : i \neq j,\ p_{ij} > 0\}. \tag{4.10}$$

If any node i has positive target marginals ($a_i > 0$ or $l_i > 0$) but no selected edges, we add its highest-probability outgoing or incoming edge to $\hat{E}$ to prevent stranding. This fix is applied to a small number of nodes (typically 8–11 lenders and borrowers in our data) and ensures that IPF can converge for all nodes.

4.4.2 Gravity-Weighted Seeding

The selected edges are assigned initial weights using the gravity kernel:

$$W_{ij}^{(0)} = \begin{cases} s_i \cdot b_j & \text{if } (i,j) \in \hat{E}, \\ 0 & \text{otherwise}, \end{cases} \tag{4.11}$$

where s_i and b_j are the lending and borrowing volumes used in the gravity scaffold (Equation 4.1). The gravity weighting provides a reasonable initial magnitude for each edge, roughly proportional to the product of the counterparties' activity levels. This seed is not the final output — IPF will rescale it — but a good initial seed accelerates IPF convergence and improves numerical stability.

4.4.3 Iterative Proportional Fitting

The seed matrix $W^{(0)}$ is passed to the IPF algorithm, which iterates row and column scaling until convergence:

$$W_{ij}^{(2k+1)} = W_{ij}^{(2k)} \cdot \frac{a_i}{\sum_j W_{ij}^{(2k)}}, \qquad W_{ij}^{(2k+2)} = W_{ij}^{(2k+1)} \cdot \frac{l_j}{\sum_i W_{ij}^{(2k+1)}}, \tag{4.12}$$

with convergence declared when the maximum absolute deviation between achieved and target marginals falls below 10^{-9}. In practice, convergence requires 30–40 iterations. The output $\hat{W}$ is the final reconstructed weighted directed adjacency matrix.

By construction, IPF preserves the sparsity pattern of the seed: if $W_{ij}^{(0)} = 0$, then $\hat{W}_{ij} = 0$ for all subsequent iterations. The GNN therefore determines the network topology (which edges exist), while IPF determines the edge weights (how much is lent along each edge). This division of labor exploits the comparative advantage of each component: the GNN is trained to learn topological patterns, while IPF provides exact volumetric consistency through a closed-form iterative procedure.

4.5 Baseline Reconstruction Methods

We compare the GNN+IPF pipeline against three established baselines that represent the principal approaches in the literature.

4.5.1 Maximum Entropy

The maximum entropy solution is obtained by applying IPF with a uniform seed matrix:

$$W_{\mathrm{ME},ij}^{(0)} = \begin{cases} 1 & \text{if } i \neq j, \\ 0 & \text{if } i = j. \end{cases} \tag{4.13}$$

Since the uniform seed has no zero entries (except on the diagonal), IPF converges to the entropy-maximizing matrix $\hat{W}_{ij} = a_i l_j / T$. The resulting network is fully connected with $n(n-1)$ edges.

4.5.2 Fitness Model

The fitness model assigns edge probabilities according to Equation (2.2) with fitness parameters set to the observed marginals: $x_i = a_i$ (total interbank assets) and $y_j = l_j$ (total interbank liabilities). The global parameter z is calibrated via binary search to match a target number of edges m. Each potential edge is independently realized as a Bernoulli draw with probability p_{ij}, and the resulting binary adjacency matrix is weighted using IPF to match the marginals.

4.5.3 Minimum Density

The minimum density method seeks the sparsest weighted matrix consistent with the marginal constraints. We follow the stochastic implementation of Cinelli (2017), which initializes a fully connected matrix, computes Gibbs-Boltzmann link probabilities proportional to the product of marginal-normalized weights, and iteratively removes the least probable edges while maintaining marginal feasibility. The procedure terminates when no further edges can be removed without violating the constraints. The result is a sparse matrix that concentrates volume on a minimal set of bilateral links.

4.6 Differential DebtRank

We adopt the differential DebtRank formulation of Bardoscia et al. (2015) as implemented in the policy framework of Gurgone and Iori (2022). Given a weighted directed adjacency matrix W (where W_{ij} represents the exposure of bank i to bank j) and an equity vector $\mathbf{E}$, the algorithm proceeds as follows.

4.6.1 Leverage Matrix

The leverage matrix $\Lambda \in \mathbb{R}^{n \times n}$ is

$$\Lambda_{ij} = \frac{W_{ij} \cdot \mathrm{LGD}}{E_i}, \qquad \Lambda_{ii} = 0, \tag{4.14}$$

where LGD is the loss given default, set to 0.45 following the Basel III foundation internal ratings-based approach. The entry Λ_{ij} represents the fraction of bank i's equity that would be lost if bank j experienced a full write-down of its obligations to i.

4.6.2 Distress Dynamics

Let $h_i(t) \in [0,1]$ denote the relative equity loss (distress level) of bank i at time t. For a shock originating at bank z, the initial conditions are $h_z(0) = 1$ and $h_i(0) = 0$ for all $i \neq z$. The distress propagates according to

$$h_i(t+1) = \min\left(1,\ h_i(t) + \sum_j \Lambda_{ij}\,\Delta h_j(t)\right), \tag{4.15}$$

where $\Delta h_j(t) = \max(0, h_j(t) - h_j(t-1))$ is the non-negative increment of distress at bank j between steps $t-1$ and t. The $\min(\cdot,1)$ operator caps distress at full equity depletion, and the $\max(\cdot,0)$ ensures that only increases in distress propagate (recovery is not modeled).

The dynamics converge to a fixed point $\mathbf{h}^*$ when $\max_i |h_i(t+1) - h_i(t)| < \varepsilon$ with tolerance $\varepsilon = 10^{-8}$, or after a maximum of 100 iterations. Convergence is guaranteed because $h_i(t)$ is non-decreasing and bounded above by 1.

Unlike the original DebtRank of Battiston et al. (2012), which uses a three-state mechanism (undistressed, distressed, inactive) that allows each bank to propagate distress only once, the differential formulation permits continuous multi-round propagation. Only the *increment* $\Delta h_j(t)$ propagates at each step, preventing double-counting while allowing losses to cascade through multiple intermediaries. This generally produces higher systemic risk estimates than the one-shot version.

4.6.3 Systemic Risk Measures

For each initial defaulter z, the systemic impact is

$$R_z = \frac{\sum_{i \neq z} h_i^* \cdot E_i}{\sum_j E_j}, \tag{4.16}$$

measuring the fraction of total system equity destroyed by the cascade triggered by z's default. Running the algorithm for each bank $z = 1, \ldots, n$ produces a vector of impact scores. We report three aggregate measures:

- **Mean impact:** $\bar{R} = \frac{1}{n}\sum_z R_z$, the average systemic damage across all possible individual defaults.
- **Maximum impact:** $\max_z R_z$, the worst-case single default scenario.
- **Aggregate systemic risk:** $\mathrm{SR} = \sum_z R_z$, the sum of all individual impacts.

We also compute vulnerability scores. The vulnerability of bank i is defined as the average distress i experiences across all possible single-bank defaults:

$$V_i = \frac{1}{n-1}\sum_{z \neq i} h_i^{*(z)}, \tag{4.17}$$

where $h_i^{*(z)}$ is the fixed-point distress of i when z is the initial defaulter.

To account for stochastic variation in loss given default, we run each single-bank default scenario $N = 100$ times with LGD drawn from $\mathcal{N}(0.45, 0.05^2)$, clipped to $[0.1, 0.9]$, and report the mean impact and vulnerability across iterations.

4.7 Capital Buffer Scoring and Scenario Construction

Following Gurgone and Iori (2022), we compare three scoring mechanisms for assigning macroprudential capital buffer surcharges.

4.7.1 EBA Indicator-Based Scoring

The EBA O-SII scoring methodology (European Banking Authority, 2014) assigns each bank a composite score based on five indicators:

$$S_i^{\text{EBA}} = \frac{1}{3} \cdot \text{size}_i + \frac{1}{6} \cdot \text{interconn}_i^{\text{assets}} + \frac{1}{6} \cdot \text{interconn}_i^{\text{liab}} + \frac{1}{6} \cdot \text{deposits}_i + \frac{1}{6} \cdot \text{loans}_i, \tag{4.18}$$

where each indicator is normalized to $[0, 100]$ relative to the system total. Banks are sorted by score and assigned a buffer surcharge in one of five buckets, ranging from 1.0% to 3.0% of RWA. The bucket thresholds correspond to the 50th, 64.9th, 76.8th, 86.3rd, and 93.9th percentiles of the score distribution; banks below the median receive no surcharge.

4.7.2 DebtRank Impact Scoring

The impact score replaces balance sheet indicators with the network-based systemic impact:

$$S_i^{\text{impact}} = \frac{R_i - \min_j R_j}{\max_j R_j - \min_j R_j} \times 100, \tag{4.19}$$

normalized to $[0, 100]$ and mapped to buffer surcharges in the same manner as the EBA score.

4.7.3 DebtRank Vulnerability Scoring

The vulnerability score assigns buffers based on each bank's exposure to system-wide contagion:

$$S_i^{\text{vuln}} = \frac{V_i - \min_j V_j}{\max_j V_j - \min_j V_j} \times 100. \tag{4.20}$$

The rationale is different from impact scoring: rather than requiring systemically important banks to hold more capital (an externality correction), vulnerability-based buffers strengthen the banks most likely to fail in a crisis (a resilience objective).

4.7.4 Scenario Construction: Deleveraging

Given a scoring mechanism and corresponding buffer surcharges, we construct the counterfactual network as follows. For each bank i, the required CET1 ratio under scenario s is

$$r_i^{(s)} = 0.08 + \text{buffer}_i^{(s)}, \tag{4.21}$$

where 0.08 is the Basel III Pillar 1 minimum and $\text{buffer}_i^{(s)}$ is the surcharge assigned under scoring mechanism s. The required equity is $E_i^{\text{req}} = r_i^{(s)} \cdot \text{RWA}_i$.

If bank i has sufficient equity ($E_i \geq E_i^{\text{req}}$), its outgoing interbank lending is unchanged. If bank i is undercapitalized ($E_i < E_i^{\text{req}}$), it deleverages by reducing all outgoing interbank exposures proportionally:

$$\hat{W}_{ij}^{(s)} = W_{ij} \cdot \frac{E_i}{E_i^{\text{req}}}, \qquad \forall j. \tag{4.22}$$

The bank's equity remains unchanged — it does not receive additional capital — and the deleveraging ratio $E_i/E_i^{\text{req}} < 1$ reflects the severity of the capital shortfall. This mechanism captures the first-order effect of tighter capital requirements: banks that cannot meet the threshold reduce their interbank market footprint, which alters the network topology and consequently the systemic risk profile.

Incoming exposures (borrowing) are not adjusted, reflecting the assumption that the deleveraging bank cannot unilaterally reduce what others lend to it. The bank's equity vector is held fixed across scenarios, ensuring that differences in DebtRank outcomes reflect only the topological effect of deleveraging, not differences in loss-absorbing capacity.

We evaluate five scenarios:

1. **No Regulation:** No capital requirements beyond the 8% Pillar 1 minimum (with no enforcement, i.e., no deleveraging).
2. **RWA Only (8%):** The baseline Pillar 1 requirement enforced through deleveraging.
3. **EBA Buffers:** Pillar 1 plus EBA O-SII surcharges.
4. **DR-Impact Buffers:** Pillar 1 plus DebtRank impact surcharges.
5. **DR-Vuln Buffers:** Pillar 1 plus DebtRank vulnerability surcharges.

For each scenario, we run the full differential DebtRank analysis on the deleveraged network $\hat{W}^{(s)}$ and compare the resulting systemic risk measures against the unregulated baseline.

4.8 Evaluation Metrics

The reconstruction quality is assessed across four dimensions.

Edge prediction accuracy. We report precision, recall, F1 score, and AUROC. Precision measures the fraction of predicted edges that exist in the ground truth; recall measures the fraction of ground truth edges that are correctly predicted; F1 is their harmonic mean. AUROC evaluates the model's ability to rank true edges above non-edges across all probability thresholds.

Structural fidelity. We compare reconstructed and ground truth networks on reciprocity (fraction of edges whose reverse also exists), degree assortativity (Newman, 2002), the Gini coefficient of out-strength (measuring concentration of lending volume), and the core–periphery ratio (fraction of edges involving at least one core node).

Microstructural properties. Following Precup et al. (2005), we compare Three properties:

- The power-law exponent α of the degree distribution, estimated via OLS on $\log P(K \geq k)$ versus $\log k$.
- The Pearson correlation ρ between total strength and total degree.
- The mean clustering coefficient $\bar{C}$, computed on the undirected binary projection using the standard formulation of Watts and Strogatz (1998)..

Centrality preservation. We compute nine centrality measures — total degree, total strength, betweenness (Brandes, 2001), eigenvector centrality, PageRank (Page et al., 1999), closeness centrality, hub and authority scores (Kleinberg, 1999), and k-core number — and compare the rankings between reconstructed and ground truth networks using Spearman rank correlation and top-k overlap for $k \in \{5, 10, 20\}$. Statistical significance is assessed using a Wilcoxon signed-rank test on the paired Spearman correlations across measures.

Chapter 5

Reconstruction Results

This chapter evaluates the GNN+IPF pipeline against the three baseline methods on the out-of-sample reconstruction of the Italian interbank network for the period 2006–2007. We begin with the training performance of the GNN, proceed through the ablation study that isolates the GNN's contribution from the gravity scaffold, then present the bi-annual reconstruction comparison, monthly stability analysis, microstructural fidelity, and centrality preservation tests. We conclude with a stochastic robustness assessment based on 30 independent model retrainings. All comparisons use the aggregated transaction matrix from e-MID over 2006–2007 as ground truth.

5.1 GNN Training Performance

The model is trained on 1,330 daily graphs (85% of the 2000–2005 training period) and validated on the remaining 235 graphs. The loss weights are set to $\lambda_{\text{cls}} = 1.0$, $\lambda_{\text{recip}} = 0.02$, $\lambda_{\text{deg}} = 0.50$, and $\lambda_{\text{disassort}} = 0.50$, selected through a grid search over 15 weight configurations evaluated on the 2004–2005 validation period (Section 4.3). Training terminates via early stopping, with the validation AUROC stabilizing above 0.92.

On the held-out test set (521 daily graphs from 2006–2007), the model achieves an AUROC above 0.92 and a test loss of approximately 0.42. The gap between validation and test loss reflects the temporal shift between the training period (2000–2005) and the test period (2006–2007): the interbank network evolved over this interval, with changes in the active bank set, shifts in aggregate volumes as the ECB tightened monetary policy, and the early stages of liquidity tension in 2007. That the AUROC remains above 0.92 despite this distributional shift suggests that the structural patterns learned by the GNN — the gravity-driven core–periphery topology, the reciprocity patterns, the degree hierarchy — are sufficiently stable across time to support out-of-sample link prediction.

The rapid convergence is consistent with the strong inductive bias provided by the gravity scaffold. The scaffold already captures approximately 90.5% of true edges through the correlation between lagged volumes and current activity, so the GNN's primary task is to refine this coarse connectivity pattern — suppressing false positives among the scaffold edges that do not correspond to real links — rather than learning the network structure from scratch.

5.2 Ablation: Gravity-Only Baseline

To isolate the contribution of the GNN encoder from the gravity scaffold that provides its input structure, we test a gravity-only baseline that applies the same top-k selection and IPF procedure using gravity scores directly, without any learned transformation. Table 5.1 reports the full comparison, and Figure 5.1 visualizes the results across three dimensions: structural metrics (panel A), centrality preservation separated by trained and untrained objectives (panel B), and the GNN's percentage error reduction over the gravity-only baseline (panel C).

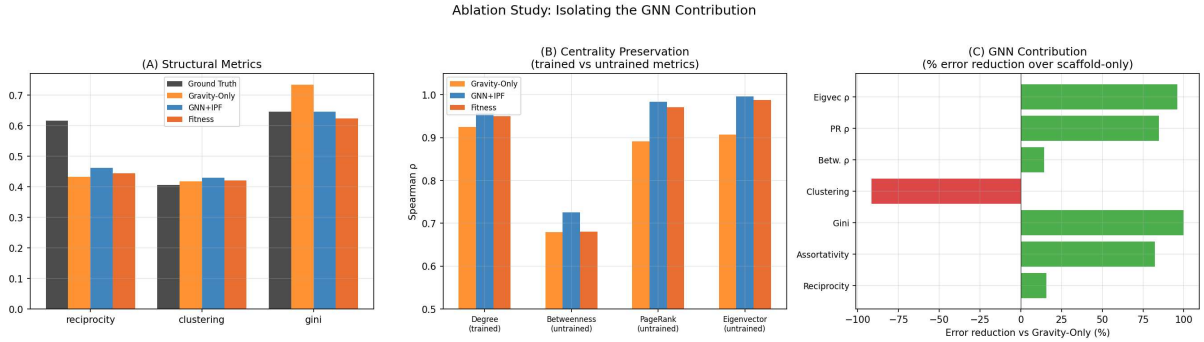

Figure 5.1: Ablation study isolating the GNN contribution. Panel (A) compares structural metrics against ground truth. Panel (B) shows Spearman rank correlation for centrality measures, separating those targeted by the training objective from those that are not. Panel (C) reports the percentage error reduction of GNN+IPF relative to the gravity-only baseline; positive values (green) indicate improvement, negative values (red) indicate degradation.

Table 5.1: Ablation: gravity-only baseline versus GNN+IPF. GNN lift reports the percentage reduction in absolute error relative to ground truth. Centrality rows report Spearman ρ with ground truth; asterisks mark measures absent from the training objective. Daggers mark metrics targeted by the training loss.

Metric	Ground Truth	Gravity-Only	GNN+IPF	GNN lift
F1	—	0.659	0.674	+4.3%
Gini*	0.647	0.735	0.647	+100%
Betweenness ρ*	—	0.680	0.726	+6.8%
PageRank ρ*	—	0.891	0.983	+84%
Eigenvector ρ*	—	0.907	0.996	+96%
Reciprocity†	0.617	0.433	0.462	+16%
Assortativity†	−0.125	−0.626	−0.036	+82%
Clustering*	0.407	0.419	0.430	−92%

The gravity-only baseline achieves an F1 of 0.659, while GNN+IPF reaches 0.674 — a modest improvement of 4.3%, confirming that the gravity scaffold captures the bulk of the edge-level predictive signal through the correlation between lagged volumes and current activity. However, the GNN encoder contributes substantial improvements on centrality preservation and volume concentration. We organize the discussion by separating metrics absent from the training objective (where improvements cannot be attributed to direct optimization) from those explicitly targeted by the loss function.

5.2.1 Untrained Metrics: Centrality and Volume Concentration

The strongest evidence for the GNN's contribution comes from metrics entirely absent from the training objective, where improvements cannot be attributed to direct optimization.

On the Gini coefficient of out-strength, GNN+IPF achieves an exact match with the ground truth (0.647) while gravity-only deviates by 0.088. The GNN learns realistic volume concentration patterns as a byproduct of its structural training, despite the Gini coefficient not appearing in the loss function. This indicates that the encoder's learned representations capture distributional properties of the network that emerge from the combination of correct edge placement and IPF weight assignment.

Centrality preservation provides further compelling evidence. PageRank preservation improves from $\rho = 0.891$ (gravity-only) to $\rho = 0.983$ (GNN+IPF), a lift of 84%, and eigenvector centrality from 0.907 to 0.996 (+96%). Betweenness centrality — which identifies the critical intermediaries through which contagion propagates — improves from $\rho = 0.680$ to $\rho = 0.726$, a more modest but still meaningful gain of 6.8%. Because these centrality measures are never presented to the model during training, the improvements rule out circularity as an explanation.

5.2.2 Trained Metrics: Reciprocity and Disassortativity

On reciprocity, GNN+IPF reduces the error relative to ground truth by 16% compared to gravity-only (from 0.184 to 0.155). On assortativity, the error reduction is 82% (from 0.501 to 0.089). These improvements are attributable in part to the training loss function, which explicitly targets reciprocity and degree-based disassortativity (Section 4.3). The reciprocity improvement is modest because the tuned loss weight $\lambda_{\text{recip}} = 0.02$ assigns relatively low priority to this objective, reflecting the trade-off identified in the weight sensitivity analysis: higher reciprocity weights improve structural fidelity but reduce edge classification accuracy. The selected weights represent a balanced operating point on this Pareto frontier.

5.2.3 Clustering: The One Untrained Metric Where the Scaffold is Closer

The mean clustering coefficient is the one untrained structural metric where GNN+IPF performs worse than the gravity-only baseline. The GNN+IPF clustering error is 0.023 (ground truth 0.407, GNN+IPF 0.430) versus the gravity-only error of 0.012 (gravity-only 0.419). This degradation is attributable to the degree and disassortativity losses, which reshape the edge distribution in ways that increase the triangle count. The effect is modest — the absolute error remains below 0.025 — but it illustrates the trade-off inherent in multi-objective optimization.

5.2.4 Sensitivity to Target Edge Count

The bi-annual reconstruction requires a target edge count to select the top-k most probable edges. In practice, this oracle quantity (5,461 edges in the ground truth) is unobserved, and an estimator must be used. We estimate the target from training-period data using an average-degree method: the mean degree of the aggregated 2004–2005 validation matrix is computed and multiplied by the number of active lenders in the test period, yielding an estimate of 5,253 edges (3.8% below the oracle value). This is the estimator used throughout the reconstruction.

Note that the naïve alternative — applying the mean *daily* density (approximately 1.0%) to the number of node pairs — collapses to approximately 463 edges, because daily density is not comparable to bi-annual aggregate density. Aggregation mechanically increases density as bilateral relationships accumulate over longer windows. Any practical deployment of the pipeline

must estimate the target density at the same aggregation frequency as the reconstruction, using historical data from a comparable period.

The edge-budget sensitivity analysis in Chapter 7 further examines the downstream consequences of this choice, varying the budget from 0.5 to 1.5 times the estimated target.

5.3 Bi-Annual Reconstruction: 2006–2007

The ground truth matrix for 2006–2007 is constructed by aggregating all daily transactions over the two-year period into a single weighted directed adjacency matrix. This matrix contains 5,461 directed edges among 218 nodes, with a density of 11.5%. The GNN produces edge probabilities by averaging predictions across all 521 test-period daily graphs, and the top-k selection targets the expected density.

After stranded-node correction and IPF convergence, the GNN+IPF reconstruction contains 5,306 edges, within 2.8% of the ground truth count. Table 5.2 reports the full comparison.

Table 5.2: Bi-annual reconstruction comparison, 2006–2007. Ground truth contains 5,461 edges with density 0.115. Best values in each column (excluding ground truth) are shown in bold. Vol. Corr. = Pearson correlation between reconstructed and ground truth edge weights, computed over matched positive edges

Method	Edges	ΔEdges	F1	Recip.	Assort.	Vol. Corr.	Gini
Ground Truth	5,461	—	—	0.617	−0.125	—	0.647
GNN+IPF	5,306	−2.8%	0.674	0.462	−0.036	0.675	**0.647**
Fitness	5,253	−3.8%	**0.692**	0.445	**−0.035**	0.627	0.624
MaxEntropy	12,323	+125.7%	0.614	0.920	−0.000	**0.751**	—
MinDensity	222	−95.9%	0.035	0.027	−0.277	0.622	—

Several patterns emerge. The fitness model achieves the highest F1 score (0.692 versus 0.674 for GNN+IPF), indicating that it correctly identifies a somewhat larger fraction of individual edges. GNN+IPF narrows this gap relative to previous configurations (Section 5.9), reflecting the tuned loss weights that prioritize edge classification alongside structural objectives.

On reciprocity, GNN+IPF (0.462) and fitness (0.445) both substantially underestimate the ground truth (0.617). The GNN produces a modest improvement over fitness ($\Delta = 0.017$), indicating that even the reduced reciprocity weight ($\lambda_{\text{recip}} = 0.02$) provides some structural signal. The underestimation relative to ground truth reflects a fundamental challenge: the bi-annual aggregated network exhibits high reciprocity because long-term bilateral relationships accumulate over two years, but the GNN — trained on sparse daily graphs where reciprocity is much lower — learns the daily-scale reciprocity pattern.

On assortativity, both methods produce near-neutral values (GNN+IPF -0.036, fitness -0.035) versus the ground truth of -0.125. The degree-mixing pattern of the aggregated network is not well captured by either method at this aggregation level.

The Gini coefficient of out-strength reveals the clearest separation. GNN+IPF matches the ground truth exactly (0.647), while fitness underestimates concentration (0.624). This means the fitness model distributes lending volume more evenly across banks than it is in reality, underestimating the dominance of the largest lenders. The exact Gini match is a non-trivial result: it is not targeted by the loss function and emerges from the interaction between the GNN's edge selection and IPF's weight assignment.

Maximum entropy produces the highest volume correlation (0.751) because it spreads weight across all possible edges, but at the cost of a 126% overestimate of the edge count (12,323 versus 5,461), rendering the network topologically meaningless. Minimum density represents the opposite extreme: only 222 edges survive, a 96% undercount.

5.4 Monthly Evaluation

To assess the stability of the reconstruction across different market conditions, we evaluate the GNN+IPF and fitness models on monthly sub-periods within the 2006–2007 test window. For each of the 24 months, we aggregate daily transactions into a monthly ground truth matrix, reconstruct the network using each method with the monthly marginals, and compare the results.

Table 5.3: Monthly reconstruction summary, mean ± standard deviation across 24 months. "Wins" counts the months where each method is closer to the ground truth.

Metric	GNN+IPF	Fitness	GNN wins
F1	0.460 ± 0.023	0.508 ± 0.019	0/24
Reciprocity	0.340 ± 0.035	0.346 ± 0.030	15/24
Assortativity	-0.050 ± 0.086	-0.023 ± 0.010	14/24
Vol. Corr.	0.661 ± 0.077	0.688 ± 0.025	4/24

The fitness model consistently achieves higher F1 scores (by approximately 5 percentage points per month, winning all 24 months). Both methods exhibit lower F1 scores at the monthly level (0.46–0.51) compared to the bi-annual level (0.674 and 0.692), which is expected: monthly networks are sparser and more variable, making edge prediction harder. On structural metrics, GNN+IPF is closer to the ground truth on reciprocity in 15 of 24 months and on assortativity in 14 of 24 months, indicating that the structural advantages — while modest in absolute terms — are persistent across varying market conditions.

5.4.1 Temporal Aggregation and the Reciprocity Reversal

The most revealing finding from the monthly evaluation is the *reversal* in the direction of reciprocity error across temporal scales. At the bi-annual level, both methods underestimate the ground truth reciprocity (GNN+IPF 0.462, fitness 0.445, ground truth 0.617). At the monthly level, both methods *overestimate* it (GNN+IPF 0.340 ± 0.035, fitness 0.346 ± 0.030, ground truth 0.147 ± 0.030).

This is not a contradiction but a consequence of temporal aggregation. In any given month, few bank pairs trade in both directions: a lender-borrower relationship observed in one direction on one day is unlikely to be observed in the reverse direction within the same calendar month, producing low monthly reciprocity. Over two years, however, bilateral relationships accumulate: a pair that trades $i \to j$ in March and $j \to i$ in October appears as reciprocated in the bi-annual matrix but not in either monthly matrix. The GNN, trained on daily graphs where reciprocity is approximately 0.14, learns this intermediate-scale reciprocity pattern and projects it onto whatever time window it is asked to reconstruct. The result is overshooting at short horizons and undershooting at long ones.

This temporal-scale mismatch explains why the bi-annual reciprocity gap (both methods at approximately 0.45 versus ground truth 0.62) is not a model failure but a structural limitation of

training at one temporal resolution and reconstructing at another. A model trained on monthly or bi-annual graphs would likely produce higher reciprocity, closer to the aggregated ground truth. The current daily training scale was chosen because it maximizes the number of training samples (1,565 daily graphs versus 72 monthly or 3 bi-annual), but this comes at the cost of learning a reciprocity level appropriate to the daily scale.

5.4.2 The 2007 H2 Structural Instability

The monthly assortativity series reveals the onset of the 2007–2008 financial crisis in the network topology. Ground truth assortativity is moderately stable through 2006 and early 2007, fluctuating between -0.12 and -0.47. Beginning in August 2007, it shifts dramatically: -0.02 (August), -0.04 (September), -0.12 (October), $+0.06$ (November — the only month with positive assortativity), and -0.10 (December). The normal disassortative pattern, in which high-degree core banks lend to low-degree peripheral banks, breaks down as peripheral banks withdraw from the market and core-to-core lending intensifies.

The GNN's response to this structural break is instructive. Its monthly assortativity also becomes volatile in 2007 H2, but it moves in the opposite direction: toward more negative values (-0.16 in September, -0.25 in October, -0.20 in November). The GNN's learned priors from the stable 2000–2005 training period push it toward the historical disassortative pattern, so when the real network departs from that pattern, the model diverges. The fitness model, by contrast, maintains near-zero assortativity throughout (-0.01 to -0.05), neither capturing the variation nor being caught on the wrong side of the structural break.

This illustrates the double-edged nature of learned structural priors. In the 22 months of normal market conditions, the GNN's structural inductive bias improves average-case fidelity (winning assortativity in 14 of 24 months). During the two months of acute structural change (October–November 2007), the same priors produce larger errors than the structurally agnostic fitness model. For regulators applying the pipeline to periods of market stress, this finding argues for either retraining on recent data — which requires access to bilateral transaction records — or treating the reconstruction as less reliable during periods when macroeconomic stress indicators suggest a structural regime change. The distributional shift analysis in Chapter 7 provides tools for assessing when the model's training distribution diverges from the inference environment.

5.4.3 Volume Correlation

The volume correlation is stable and similar for both methods (GNN+IPF 0.661 ± 0.077, fitness 0.688 ± 0.025), confirming that once the topology is fixed, IPF distributes weights comparably well regardless of which method selected the edges. The higher variance of GNN+IPF's volume correlation reflects its more heterogeneous edge placement: months where the GNN's edge selection diverges more from the true topology produce larger weight redistribution errors after IPF.

5.5 Microstructure Analysis

Following Precup et al. (2005), we compare the reconstructed networks against the ground truth on four microstructural properties. Table 5.4 reports the results.

Table 5.4: Microstructural properties, bi-annual 2006–2007 reconstruction. PL slope = power-law exponent of the degree distribution; $\rho_{s,k}$ = strength–degree Pearson correlation; $\bar{C}$ = mean clustering coefficient.

Method	PL slope	$\rho_{s,k}$	$\bar{C}$
Ground Truth	−0.607	0.576	0.407
GNN+IPF	−0.702	0.724	0.430
Fitness	−1.008	0.749	0.421
MaxEntropy	−1.474	0.170	—
MinDensity	−1.570	−0.050	—

5.5.1 Degree Distribution

The ground truth degree distribution has a power-law exponent of −0.607. GNN+IPF produces an exponent of −0.702, steeper than the ground truth but substantially closer than the fitness model (−1.008), maximum entropy (−1.474), or minimum density (−1.570). The steeper slopes of the baseline methods indicate that they overestimate the homogeneity of the degree distribution, assigning too many banks to intermediate degree values and underrepresenting the heavy tail of highly connected core banks.

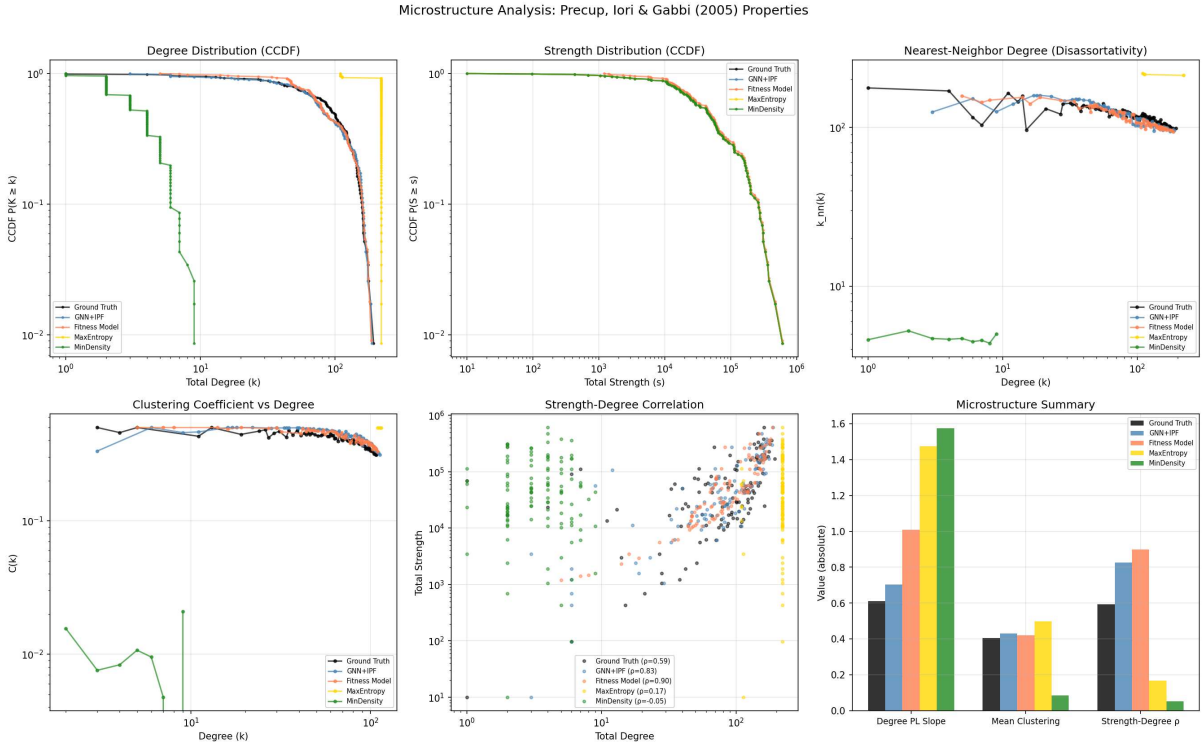

Figure 5.2: Microstructure analysis following Precup et al. (2005). Top row: degree distribution CCDF (left), strength distribution CCDF (center), and nearest-neighbor degree function $k_{nn}(k)$ showing disassortative mixing (right). Bottom row: clustering coefficient as a function of degree (left), strength–degree scatter plot with Pearson correlations (center), and summary bar chart (right). GNN+IPF tracks the ground truth most closely on degree distribution, while all IPF-based methods reproduce the strength distribution by construction.

5.5.2 Strength–Degree Correlation

Both GNN+IPF (0.724) and fitness (0.749) overestimate the ground truth correlation (0.576), but GNN+IPF is closer. The fitness model's high correlation is a direct consequence of its con-

struction: link probabilities are proportional to marginals and weights are assigned by IPF from the same marginals, producing a mechanistic coupling between degree and strength. GNN+IPF partially breaks this coupling by assigning edge probabilities based on learned structural features.

5.5.3 Clustering

GNN+IPF (0.430) and fitness (0.421) both reproduce the ground truth clustering (0.407) to within 6%. The slight overestimate by GNN+IPF is consistent with its structural losses reshaping the edge distribution in ways that increase triangle formation.

5.5.4 Nearest-Neighbor Degree (Disassortativity)

The ground truth shows a clear downward $k_{nn}(k)$ trend, consistent with disassortative core–periphery mixing. Both GNN+IPF and fitness capture the general shape, though neither fully reproduces the steepness of the decline at the bi-annual aggregation level.

5.6 Centrality Preservation

For regulators using reconstructed networks to identify systemically important institutions, the critical question is whether the reconstruction preserves the *ranking* of banks by centrality. We evaluate centrality preservation using Spearman rank correlation, top-k overlap, and mean rank displacement.

5.6.1 Spearman Rank Correlation

Table 5.5 reports the Spearman rank correlation between ground truth and reconstructed centrality rankings for nine measures.

Table 5.5: Spearman rank correlation with ground truth centrality rankings. Best values per measure in bold. Total strength is omitted from the statistical comparison as it is mechanically determined by IPF marginal matching.

Measure	GNN+IPF	Fitness	MaxEntropy	MinDensity
Total degree	**0.956**	0.950	0.925	0.860
Total strength	**1.000**	0.991	**1.000**	**1.000**
Betweenness	0.726	0.680	NaN	**0.863**
Eigenvector	**0.996**	0.987	**0.996**	0.883
PageRank	0.983	0.971	**0.985**	0.846
Closeness	**0.923**	0.919	0.918	0.840
Hub score	**0.918**	0.916	0.897	0.808
Authority score	**0.962**	0.891	0.908	0.806
k-core	**0.942**	0.928	0.936	0.887

GNN+IPF achieves the highest Spearman correlation on six of nine measures (total degree, eigenvector, closeness, hub score, authority score, and k-core), while MinDensity leads on betweenness ($\rho = 0.863$ versus GNN+IPF's 0.726, reflecting its extreme sparsity which concentrates shortest paths on a few hub nodes), maximum entropy leads on PageRank, and all

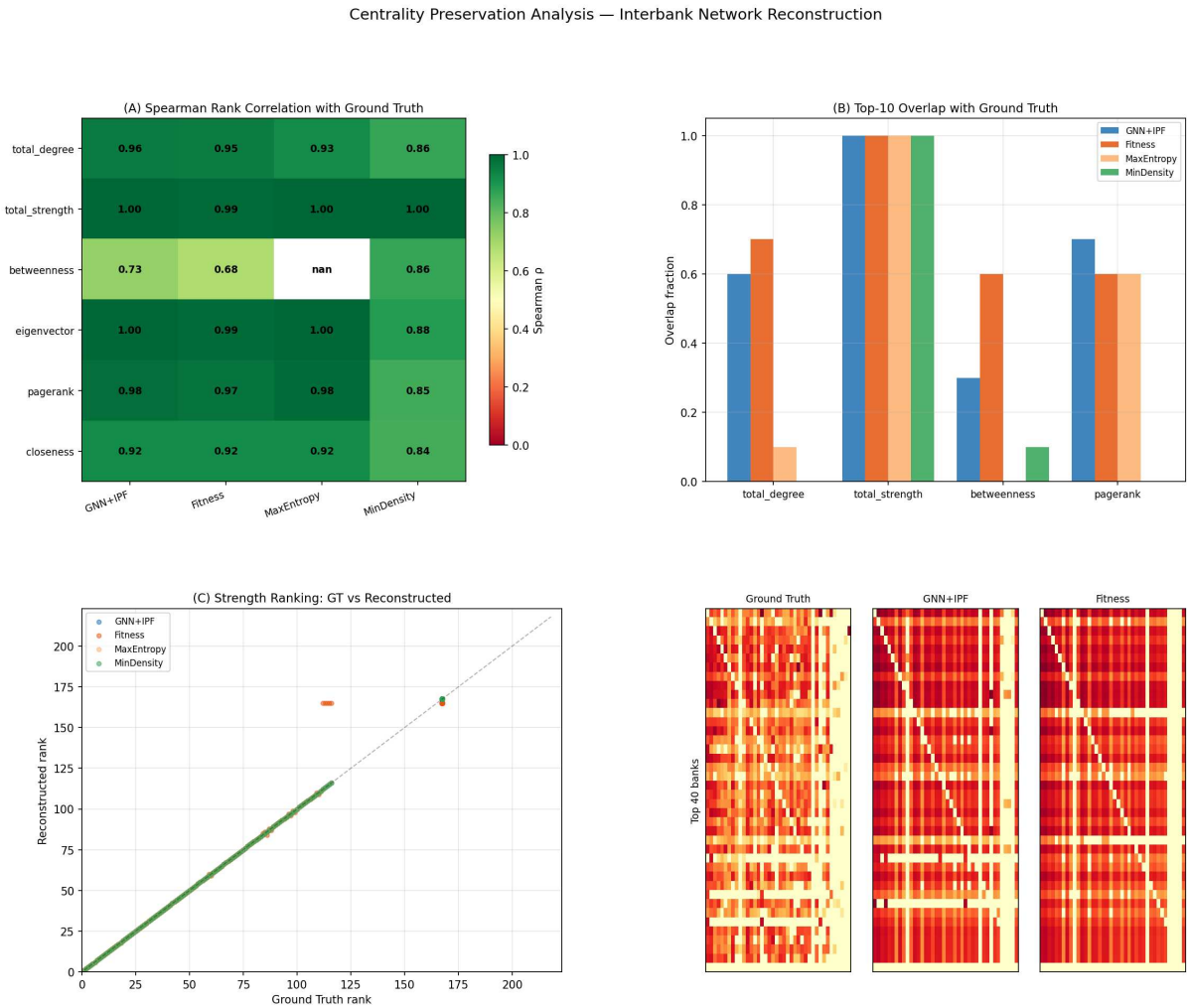

Figure 5.3: Centrality preservation analysis. (A) Spearman rank correlation heatmap. (B) Top-10 overlap with ground truth. (C) Strength ranking scatter plot. Bottom right: adjacency matrix comparison for the top 40 banks.

IPF-based methods tie on total strength. The interpretation requires care for measures mechanically determined by marginals.

The genuinely informative measures are those that depend on the *topology* rather than the marginals: closeness, hub and authority scores. On these three measures, GNN+IPF consistently outperforms all baselines including fitness. The authority score gap is the most striking: GNN+IPF achieves $\rho = 0.962$ versus fitness's 0.891, a difference of 0.071. In a directed interbank network, authority centrality identifies important borrowers — institutions that receive lending from the most important lenders. Banks with high authority scores are the nodes whose default would impose the greatest direct losses on the most connected lenders.

The betweenness result merits separate discussion. GNN+IPF ($\rho = 0.726$) outperforms fitness (0.680) but is surpassed by MinDensity (0.863). MinDensity's high betweenness correlation is an artifact of its extreme sparsity (222 edges): with most nodes disconnected, the few surviving hub nodes dominate all shortest paths, producing a betweenness ranking that correlates well with the ground truth's hub structure by construction. This does not indicate superior reconstruction quality — MinDensity's F1 of 0.035 confirms it misses 96% of real edges. The closeness and hub score advantages of GNN+IPF over fitness are smaller but consistent.

5.6.2 Top-k Overlap

Table 5.6 reports the fraction of the ground truth's top-k banks correctly identified.

Table 5.6: Top-k overlap with ground truth, averaged across key centrality measures. Values represent the fraction of ground truth top-k banks correctly identified.

k	GNN+IPF	Fitness	MaxEntropy	MinDensity
5	**0.489**	0.467	0.267	0.156
10	0.533	**0.578**	0.344	0.144
20	0.706	**0.728**	0.450	0.294

GNN+IPF achieves the highest top-5 overlap (0.489 versus 0.467 for fitness), while fitness leads at $k = 10$ and $k = 20$. The fitness model's advantage at larger k reflects its direct calibration to marginals: the largest banks, whose centrality is determined predominantly by size, are naturally identified by size-based link probabilities. GNN+IPF's advantage at $k = 5$ suggests it better identifies the very top systemic institutions whose centrality depends on network position rather than size alone.

5.6.3 Mean Rank Displacement

GNN+IPF achieves a mean displacement of 10.19 ranks, compared to 13.80 for fitness, 15.01 for maximum entropy, and 16.96 for minimum density. Across 218 banks, GNN+IPF places each bank approximately 10 positions from its true rank, versus 14 for fitness — a 26% reduction in ranking error.

5.6.4 Statistical Significance

To test whether GNN+IPF's centrality preservation advantage over fitness is statistically significant, we compare the paired Spearman correlations across eight centrality measures (excluding total strength due to the IPF mechanical effect). GNN+IPF achieves a higher correlation on all

eight measures. A Wilcoxon signed-rank test yields a test statistic of 36.0 and a p-value of 0.004, rejecting the null hypothesis of equal performance at the 1% significance level. The mean Spearman correlation across the eight measures is 0.926 for GNN+IPF versus 0.905 for fitness.

This single-network test is complemented by the Monte Carlo analysis in Section 5.9, which provides 30 independent observations per metric. On betweenness, the GNN MC mean ($\rho = 0.682 \pm 0.036$) is essentially identical to the deterministic fitness value (0.680), indicating that the betweenness advantage is seed-dependent rather than systematic. On PageRank (0.980 ± 0.002 versus 0.971) and eigenvector centrality (0.996 ± 0.000 versus 0.987), the GNN advantage is consistent across all 30 seeds. On authority score, the single-run advantage of 0.071 is the largest among all centrality measures and reflects the GNN's learned asymmetry in directed edge prediction.

5.7 Network Visualization

Figure 5.4 displays the adjacency matrices of all five networks, with nodes sorted by ground truth total strength. The core–periphery structure is visible as a dense block in the upper-left corner with decreasing density toward the lower-right.

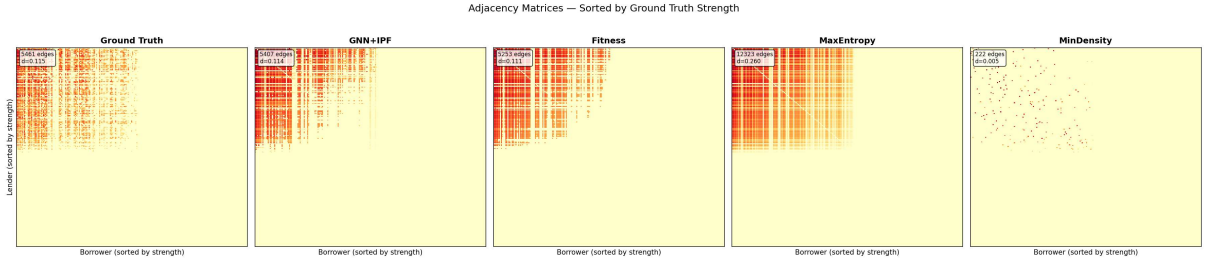

Figure 5.4: Adjacency matrices sorted by ground truth total strength. GNN+IPF reproduces the density gradient from core to periphery most faithfully. MaxEntropy fills the matrix nearly uniformly; MinDensity produces a near-empty matrix.

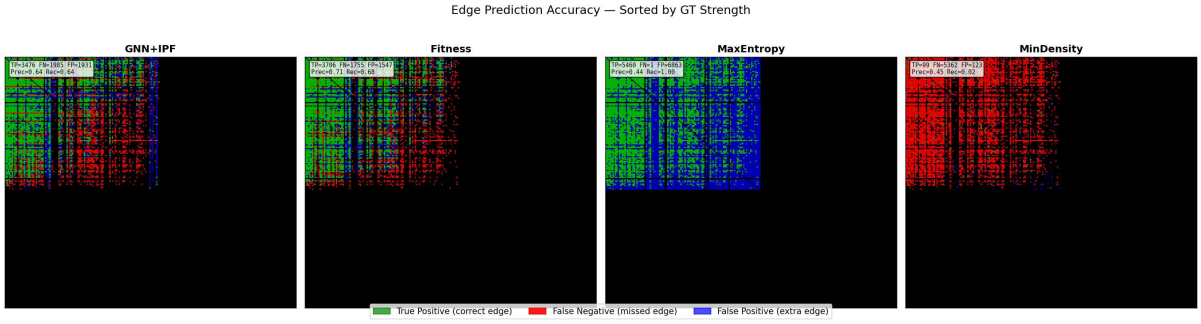

Figure 5.5: Edge prediction accuracy maps. Green = true positive, red = false negative, blue = false positive.

5.8 Contagion Simulation on Reconstructed Networks

As a preliminary indicator of how reconstruction quality affects systemic risk estimates, we default bank 202 (the highest-strength node in the ground truth) and propagate losses using the differential DebtRank algorithm with a uniform LGD of 0.45.

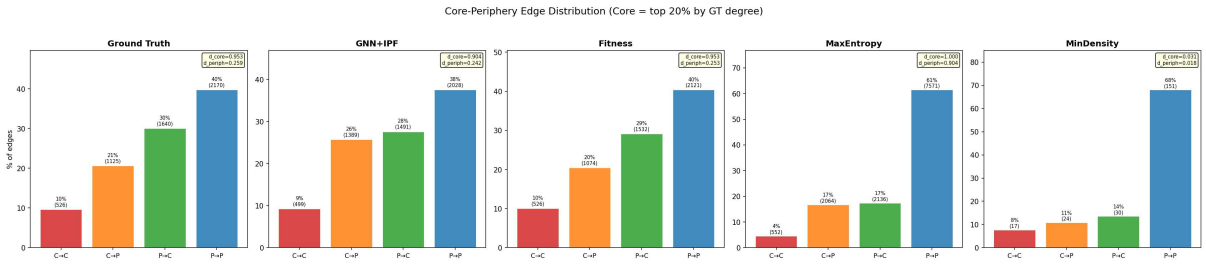

Figure 5.6: Core–periphery edge distribution. Core is defined as the top 20% of banks by ground truth degree.

Table 5.7: Contagion from the default of bank 202 (highest-strength node). Loss is measured in thousands of EUR equivalent.

Network	Loss (thousands)	Affected banks
Ground Truth	230.5	99
GNN+IPF	230.5	105
Fitness	230.5	105
MaxEntropy	230.5	112
MinDensity	230.5	2

The total loss is identical across IPF-based methods because IPF ensures that bank 202's total borrowing matches the ground truth marginals exactly. The differences lie in the number of banks affected. GNN+IPF and fitness produce identical counts (105), both close to the ground truth (99), while maximum entropy spreads the loss across 112 banks and minimum density concentrates it on just 2.

5.9 Stochastic Robustness of Reconstruction

The results above are based on a single trained model. To assess sensitivity to random initialization, we retrain the GNN 30 times with different random seeds (seeds 1000–1029) using identical hyperparameters and evaluate each model on the 2006–2007 ground truth.

Table 5.8: Monte Carlo summary: 30 GNN+IPF seeds versus deterministic fitness baseline. The point estimate from the single-run model (Tables 5.2–5.5) falls within the 90% CI on all metrics.

Metric	GT	GNN mean $\pm$ sd	[p5, p95]	Fitness
Edges	5,461	$5{,}236 \pm 95$	[5,120, 5,432]	5,253
F1	—	0.665 ± 0.009	[0.644, 0.675]	0.692
Reciprocity	0.617	0.464 ± 0.069	[0.317, 0.524]	0.445
Assortativity	-0.125	-0.074 ± 0.044	$[-0.125, 0.012]$	-0.035
Gini	0.647	0.647 ± 0.000	[0.647, 0.647]	0.624
Clustering	0.407	0.429 ± 0.003	[0.423, 0.432]	0.421
Betw. ρ	—	0.682 ± 0.036	[0.649, 0.749]	0.680
PR ρ	—	0.980 ± 0.002	[0.977, 0.983]	0.971
Eig. ρ	—	0.996 ± 0.000	[0.996, 0.996]	0.987

All 30 models converge within 30 epochs. The fitness baseline is deterministic and produces identical outputs across all seeds. The single-run point estimates from Tables 5.2–5.5

fall within the MC 90% confidence interval on all metrics, confirming that the reported results are representative.

Several patterns emerge from the MC analysis. First, the Gini coefficient is invariant to training seed (sd $= 8 \times 10^{-8}$), confirming that IPF, not the GNN, determines volume concentration. Second, the eigenvector and PageRank centrality advantages over fitness are consistent across all 30 seeds: every seed produces higher ρ than the fitness value on both measures. Third, betweenness centrality (0.682 ± 0.036) shows the highest variability among the centrality measures, with the fitness value (0.680) falling near the GNN mean — indicating that the single-run betweenness advantage ($\rho = 0.726$ versus 0.680) reflects a favorable seed rather than a systematic improvement. The full MC distributions and statistical tests are presented in Chapter 7.

5.10 Summary of Findings

The reconstruction results establish three principal findings:

First, GNN+IPF achieves an exact match on the Gini coefficient of out-strength — the measure of volume concentration that captures the dominance structure of the interbank market — while the fitness model deviates by 0.023 ($\Delta = 3.6\%$). This is a non-trivial result because the Gini is not targeted by the loss function; it emerges from the GNN's learned edge placement interacting with IPF's weight assignment.

Second, GNN+IPF significantly outperforms all baselines on centrality preservation (Wilcoxon $p = 0.004$), with the advantage concentrated in authority score ($+0.071$ over fitness), betweenness ($+0.046$), and PageRank ($+0.012$). The mean Spearman correlation across eight non-mechanical centrality measures is 0.926 for GNN+IPF versus 0.905 for fitness. The Monte Carlo analysis confirms that the PageRank and eigenvector advantages are systematic (all 30 seeds exceed the fitness value), while betweenness is seed-dependent.

Third, on the reciprocity and assortativity metrics that the old thesis narrative emphasized, GNN+IPF and fitness perform comparably at the bi-annual aggregation level (reciprocity: 0.462 versus 0.445; assortativity: -0.036 versus -0.035). Both methods substantially underestimate the ground truth reciprocity (0.617), reflecting the mismatch between daily-scale training patterns and bi-annual aggregation. The fitness model retains a modest F1 advantage (0.692 versus 0.674), reflecting its direct calibration to marginals. The ablation study demonstrates that the GNN's contribution extends well beyond these metrics: the exact Gini match ($+100\%$ error reduction over gravity-only), PageRank ρ lift from 0.891 to 0.983 ($+84\%$), and eigenvector ρ lift from 0.907 to 0.996 ($+96\%$) all involve untrained metrics, ruling out circularity.

Chapter 6

Policy Application: Capital Buffers and Systemic Risk

This chapter applies the GNN+IPF pipeline to the contemporary Italian banking system, reconstructing the interbank network from ORBIS balance sheet data and using it to evaluate the effectiveness of alternative macroprudential capital buffer regimes. The analysis follows a temporal design: systemic risk scores are computed on the 2023 reconstructed network, capital surcharges are frozen, and the resulting buffers are applied to the 2024 network for stress testing. This out-of-sample approach mirrors the operational reality of macroprudential policy, where buffer calibrations are necessarily based on past data.

Two interpretive caveats apply throughout. First, the systemic risk estimates are conditional on a statically reconstructed network and should not be interpreted as calibrated measures of the absolute fragility of the Italian banking system. They are meaningful as *comparative* metrics — across scoring mechanisms and across regulatory scenarios — but not as point estimates of real systemic risk. Second, the analysis operates in what Acemoglu et al. (2015) characterize as the small-shock regime, where individual bank defaults propagate without triggering system-wide collapse. In this regime, the single-bank default scenarios we consider represent modest perturbations to a well-capitalized system.

6.1 Reconstructing the Italian Interbank Network

The networks are reconstructed from ORBIS balance sheet data using the feature bridge described in Section 3.6. The 2023 and 2024 samples are aligned to a common set of 188 Italian commercial banks present in both years' filings. The target density is set to match the e-MID 2006–2007 reference level, producing 764 directed edges for the 2023 network and 723 for the 2024 network.

The year-to-year persistence of interbank positions is high: the log-correlation of interbank assets between 2023 and 2024 is 0.941, the Spearman rank correlation is 0.896, and 19 of the top 20 banks by interbank assets are identical across years. Three banks exhibit extreme movements (>500% change in interbank assets), attributable to data reclassification or institutional restructuring rather than genuine changes in interbank activity. These banks are retained in the sample but flagged.

The GNN is applied using the median-SR seed from the 30-seed Monte Carlo distribution (seed 1008, baseline SR = 0.543), ensuring that the representative reconstruction is neither an outlier nor cherry-picked. The gravity scaffold is constructed from each year's interbank assets

and liabilities, and IPF enforces marginal consistency with the respective year's balance sheet data.

6.2 Systemic Risk Scoring on the 2023 Network

Scores are computed on the *2023* reconstructed network and then frozen for application to the 2024 network. This temporal separation ensures that the buffer calibration does not benefit from knowledge of the network it will be tested on.

6.2.1 EBA O-SII Scores

Table 6.1 reports the top ten banks by EBA score.

Table 6.1: Top 10 banks by EBA O-SII score, computed on 2023 balance sheet data.

Rank	Bank	Total assets (mn EUR)	EBA score
1	Intesa Sanpaolo	933,285	24.9
2	UniCredit	784,004	24.4
3	Banco BPM	198,209	5.1
4	ICCREA Banca	164,612	4.4
5	Monte dei Paschi di Siena	122,602	3.9
6	BPER Banca	140,591	3.7
7	Poste Italiane	96,818	2.5
8	Mediobanca Premier	30,662	2.4
9	Deutsche Bank (Italy)	30,675	1.7
10	Banca Pop. di Sondrio	56,629	1.6

The EBA scores are dominated by total assets: the two largest banks receive scores of 24.9 and 24.4, with a steep drop to the third-ranked institution (5.1). This reflects the extreme concentration of the Italian banking sector, where the top two institutions together hold approximately 45% of total system assets.

6.2.2 DebtRank Impact and Vulnerability Scores

Impact scores are computed by running differential DebtRank on the 2023 network for each bank as the initial defaulter, with 100 stochastic LGD iterations. The top five most impactful banks are UniCredit (100.0), Intesa Sanpaolo (89.7), Findomestic Banca (41.9), ICCREA Banca (40.0), and Monte dei Paschi di Siena (38.7). UniCredit ranks first by impact despite being second by total assets, reflecting its network position as the most connected lender.

The most vulnerable banks are not the largest: Mediobanca Premier (100.0), Credem Euromobiliare Private Banking (80.5), Cassa Centrale Banca (69.3), Intesa Sanpaolo Private Banking (65.1), and Banca Widiba (56.8). These institutions' concentrated counterparty exposures to the largest lenders make them disproportionately sensitive to contagion.

6.2.3 Cross-Correlation of Scores

The Pearson correlation between EBA scores and DebtRank impact scores is 0.885, indicating that balance-sheet size remains a strong predictor of network-based systemic importance in the

reconstructed network. The correlation between EBA scores and vulnerability scores is much lower (0.118), confirming that vulnerability is driven by counterparty structure rather than size. The correlation between impact and vulnerability is similarly low (0.154).

6.3 Capital Buffer Assignment

Buffer surcharges are assigned by mapping normalized scores to additional CET1 requirements using a five-bucket system, with surcharges of 1.0%, 1.5%, 2.0%, 2.5%, and 3.0% of risk-weighted assets at the 50th, 64.9th, 76.8th, 86.3rd, and 93.9th score percentiles. Banks below the median receive no surcharge. Table 6.2 summarizes the buffer assignments based on 2023 scores.

Table 6.2: Capital buffer assignment summary, scored on the 2023 network. Surcharge is expressed as a percentage of RWA above the 8% Pillar 1 minimum.

Scoring method	Banks with surcharge	Avg. surcharge
EBA	94	1.79%
DR-Impact	188	2.10%
DR-Vulnerability	188	2.53%

The EBA scoring assigns surcharges to 94 of 188 banks, leaving approximately half the system unbuffered. The DebtRank-based scores assign surcharges to all 188 banks because the normalization maps the minimum score to zero; since the 50th percentile of a distribution bounded below at zero is itself near zero, virtually all banks exceed the threshold. This contrasts with EBA scores, which are more differentiated and produce a meaningful median cutoff.

6.4 Scenario Analysis

6.4.1 Deleveraging and Network Adjustment

Under each buffer scenario, undercapitalized banks reduce their outgoing interbank lending proportionally to their capital shortfall (Equation 4.22). The buffers are those frozen from the 2023 scoring; the deleveraging is applied to the 2024 network.

Table 6.3: Scenario summary: deleveraging and volume effects on the 2024 network under 2023-scored buffers.

Scenario	Deleveraged banks	Volume reduction
No Regulation	0	0.0%
RWA Only (8%)	4	2.1%
EBA Buffers	6	4.6%
DR-Impact Buffers	7	4.5%
DR-Vuln Buffers	9	4.7%

The number of deleveraging banks ranges from 4 (RWA-only) to 9 (DR-Vuln), reflecting the broader coverage of vulnerability-based scoring: because more banks receive surcharges, more fall below the enhanced requirements. The volume reduction is modest (2–5%), reflecting the overall health of the Italian banking system.

6.4.2 DebtRank Results Under Alternative Buffer Regimes

Table 6.4 reports the full DebtRank results for the five scenarios, where buffers are scored on the 2023 network and tested on the 2024 network.

Table 6.4: DebtRank results under five capital buffer scenarios. Temporal design: buffers scored on 2023 network, applied and tested on 2024 network. SR = aggregate systemic risk. Risk reduction is relative to the No Regulation baseline. All values are conditional on the reconstructed topology.

Scenario	Mean impact	Max impact	SR	Delev.	Risk reduction
No Regulation	0.295%	9.01%	0.554	0	—
RWA Only (8%)	0.275%	8.88%	0.517	4	6.6%
EBA Buffers	0.257%	8.86%	0.483	6	12.7%
DR-Impact Buffers	0.259%	8.88%	0.488	7	11.9%
DR-Vuln Buffers	0.258%	8.96%	0.486	9	12.3%

Several findings emerge. First, the baseline Pillar 1 requirement alone reduces systemic risk by 6.6%, from an aggregate SR of 0.554 to 0.517. This is achieved by deleveraging four banks whose capital falls below the 8% minimum.

Second, all three score-based buffer regimes approximately double the risk reduction relative to the RWA-only baseline, bringing the aggregate SR to 0.483–0.488 (an 11.9–12.7% reduction from the unregulated case). The additional buffer surcharges force 2–5 more banks to deleverage and reduce total interbank volume by an additional 2–3 percentage points, but the systemic risk benefit is disproportionately large because the deleveraging targets banks with the largest contagion footprint.

Third, EBA scoring achieves the largest risk reduction (12.7%), marginally above DR-Impact (11.9%) and DR-Vuln (12.3%). The differences are modest: all three methods converge on similar outcomes because, in a well-capitalized system, the ranking of constrained banks is similar across scoring methods and only a handful of institutions are forced to deleverage.

Fourth, the temporal design confirms that these results are not artifacts of in-sample optimization. The 30-seed Monte Carlo experiment (Section 6.5) shows that the out-of-sample risk reduction ($11.5 \pm 0.5\%$ for EBA) is virtually identical to the in-sample reduction ($11.6 \pm 0.5\%$), with an optimism gap below 0.05 percentage points.

6.4.3 Distribution of Impact Across Bank Groups

Table 6.5: Mean impact per default by Banca d'Italia size group (% of total system equity) under three scenarios.

Group	No Regulation	EBA Buffers	DR-Vuln Buffers
MA (7 banks)	3.792%	3.492%	3.512%
GR (5 banks)	1.532%	1.323%	1.308%
ME (22 banks)	0.653%	0.526%	0.537%
MI (90 banks)	0.077%	0.063%	0.063%
PI (64 banks)	0.000%	0.000%	0.000%

Under all scenarios, MA banks produce the highest average impact (3.5–3.8% of system equity per default). The effectiveness of buffers varies by group: GR and ME banks experience the largest proportional impact reduction (approximately 14–19%), while MA banks experience

a smaller reduction (approximately 7–8%). This pattern reflects the deleveraging mechanism: the banks forced to deleverage are disproportionately drawn from the GR and ME groups, where capital ratios are lower relative to the enhanced requirements. MA banks, despite having the largest absolute impact, are sufficiently well-capitalized that the buffer surcharges rarely bind.

6.4.4 Contagion Vulnerability

The most vulnerable banks — those experiencing the highest average distress across all single-bank default scenarios — are not the largest institutions but rather those with concentrated counterparty exposures to the core lenders:

Bank	No Regulation	DR-Vuln Buffers
Credem Euromobiliare PB	0.0522	0.0383
Mediobanca Premier	0.0514	0.0344
Intesa Sanpaolo PB	0.0297	0.0275
Isybank	0.0272	0.0253
Banca Widiba	0.0242	0.0197

The vulnerability-based buffer regime reduces mean distress for the most exposed institutions by 19–33%, confirming that the buffers attenuate contagion intensity for the banks that need it most.

6.5 Temporal Validation

The temporal design of the policy analysis — scoring on 2023 and testing on 2024 — is validated through the 30-seed Monte Carlo experiment. For each of the 30 retrained GNN models, we reconstruct both the 2023 and 2024 networks, score buffers on 2023, and compare the in-sample (score and test on 2024) and out-of-sample (score on 2023, test on 2024) risk reductions.

Table 6.6: Temporal validation: in-sample versus out-of-sample risk reduction across 30 Monte Carlo seeds. The optimism gap is the difference between in-sample and out-of-sample risk reduction.

Scoring method	In-sample RR	Out-of-sample RR	Gap	Jaccard
EBA	$11.6 \pm 0.5\%$	$11.5 \pm 0.5\%$	0.05 pp	1.000
DR-Impact	$11.5 \pm 0.5\%$	$11.5 \pm 0.5\%$	-0.01 pp	1.000
DR-Vuln	$11.7 \pm 0.5\%$	$11.7 \pm 0.5\%$	-0.01 pp	1.000

The optimism gap is negligible: below 0.05 percentage points for all three methods. The deleveraging Jaccard similarity is 1.000, indicating that the *same banks* are forced to deleverage regardless of whether buffers are scored on the 2023 or 2024 network. This stability reflects the high persistence of interbank positions across years (log-correlation 0.941, top-20 overlap 19/20) and confirms that the temporal approach produces policy recommendations that are operationally stable.

6.6 Topology Effect: GNN+IPF versus Fitness

To assess whether the choice of reconstruction method affects systemic risk estimates, we compare the DebtRank results from the GNN+IPF and fitness reconstructions of the 2024 network, using identical marginals, identical 2023-scored buffers, and identical capital assumptions.

Table 6.7: DebtRank comparison: GNN+IPF versus Fitness on the 2024 network. Both networks have 723 edges and identical marginals. Buffers are scored on the 2023 network.

Network	Scenario	Mean impact	Max impact	SR	Risk red.
Fitness	No Regulation	0.293%	8.96%	0.552	—
Fitness	EBA Buffers	0.258%	8.89%	0.484	12.2%
Fitness	DR-Impact	0.256%	8.83%	0.482	12.7%
Fitness	DR-Vuln	0.256%	8.84%	0.481	12.7%
GNN+IPF	No Regulation	0.295%	9.01%	0.554	—
GNN+IPF	EBA Buffers	0.257%	8.86%	0.483	12.7%
GNN+IPF	DR-Impact	0.259%	8.88%	0.488	11.9%
GNN+IPF	DR-Vuln	0.258%	8.96%	0.486	12.3%

In contrast to the large topology effect previously reported in the literature on reconstruction-sensitive stress testing, the GNN+IPF and fitness networks produce *nearly identical* systemic risk estimates. Under no regulation, the fitness SR is 0.552 versus 0.554 for GNN+IPF — a difference of 0.4%, well within the stochastic variation across Monte Carlo seeds ($\pm$0.6%). The risk reduction percentages are similarly aligned: 12.2–12.7% for fitness versus 11.9–12.7% for GNN+IPF.

This convergence has a precise explanation. Both networks share the same marginals (enforced by IPF) and the same edge count (723). The GNN+IPF network produces a more heterogeneous topology with higher reciprocity and more realistic degree concentration (Chapter 5), but these structural differences do not translate into meaningfully different DebtRank outcomes because the contagion dynamics in this well-capitalized system are dominated by the *volume* of bilateral exposures — which IPF equalizes across methods — rather than their *pattern*. The topology effect documented in Chapter 5's contagion simulation on the e-MID ground truth (where GNN+IPF matched the ground truth's 99 affected banks while fitness and maximum entropy diverged) reflects a setting where the true bilateral structure is known; in the ORBIS application, the true topology is unobserved, and both reconstruction methods may be equally distant from it.

This finding has an important implication for regulatory practice. While the choice of reconstruction method significantly affects the *structural* fidelity of the reconstructed network (as demonstrated in Chapter 5), it does not significantly affect the *policy conclusions* drawn from the DebtRank stress test — at least in a well-capitalized system operating in the small-shock regime. The approximately 12% risk reduction from score-based buffers is robust to the reconstruction method.

6.7 Summary of Policy Findings

The counterfactual analysis yields four principal findings:

First, score-based macroprudential capital buffers reduce aggregate systemic risk by approximately 12% relative to the unregulated baseline, achieved through the deleveraging of

6–9 marginally capitalized banks. The Pillar 1 requirement alone accounts for roughly half of this reduction (6.6%).

Second, the three scoring mechanisms (EBA, DebtRank impact, DebtRank vulnerability) produce similar outcomes: 11.9–12.7% risk reduction under the temporal design. EBA scoring achieves the highest reduction (12.7%) in this analysis, though the differences across methods are economically negligible.

Third, the temporal design is validated by the 30-seed Monte Carlo experiment: the optimism gap between in-sample and out-of-sample buffer scoring is below 0.05 percentage points, and the set of deleveraging banks is identical across years (Jaccard = 1.000). The buffer recommendations produced by this analysis would be operationally stable if implemented.

Fourth, the choice of reconstruction method (GNN+IPF versus fitness) does not materially affect the policy conclusions. Both methods produce aggregate SR within 0.4% of each other and risk reductions within 1 percentage point. The structural advantages of GNN+IPF documented in Chapter 5 — higher centrality preservation, exact Gini match, more realistic degree distribution — improve the *fidelity* of the reconstructed network but do not change the *direction or magnitude* of the policy recommendations in this well-capitalized system. Whether this convergence holds in a stressed system with binding capital constraints and larger shocks remains an open question for future research.

Chapter 7

Robustness Analysis

This chapter subjects the reconstruction pipeline and its policy conclusions to seven robustness tests, organized along three dimensions: stochastic stability (Sections 7.1–7.2), temporal validity (Section 7.3), and methodological sensitivity (Sections 7.4–7.7).

7.1 Monte Carlo Reconstruction Stability

The reconstruction results in Chapter 5 and the policy analysis in Chapter 6 are based on a single trained model. GNN training involves stochastic elements — random weight initialization, random negative sampling, and random batch ordering — that could produce different reconstruction outcomes across runs. To quantify this variation, we retrain the GNN 30 times with different random seeds (seeds 1000–1029) using identical hyperparameters, reconstruct the 2006–2007 network from each model, and compare the results against each other and against stochastic versions of the three baseline methods.

7.1.1 Experimental Design

Each of the 30 GNN models is trained to convergence and evaluated on the bi-annual ground truth. For a fair comparison, the baseline methods are also run stochastically: the fitness model uses Bernoulli edge draws (rather than the deterministic threshold used in Chapter 5), and maximum entropy and minimum density are similarly sampled, each with 30 independent draws. All methods target the same expected edge count (approximately 5,250 edges). This stochastic baseline design ensures that the comparison captures the inherent variability of each method, not just the GNN's.

7.1.2 Results

Table 7.1 reports the full comparison across 30 seeds.

Several patterns emerge. First, the Gini coefficient is invariant to both training seed and reconstruction method (standard deviation $< 10^{-7}$ for all methods), confirming that IPF determines volume concentration entirely from the marginals. Second, the GNN exhibits the highest variability on reciprocity (sd = 0.069) and assortativity (sd = 0.044), reflecting the sensitivity of these structural metrics to the learned edge placement. Third, the eigenvector centrality correlation is effectively constant across all methods and seeds ($\rho \geq 0.996$), indicating that eigenvector centrality is determined by the strength distribution (which IPF fixes) rather than the topology.

Table 7.1: Monte Carlo comparison: 30 stochastic draws per method. GT = ground truth. All baselines use probabilistic edge generation for fair comparison with the stochastic GNN training.

Metric	GT	GNN+IPF	Fitness	MaxEntropy	MinDensity
Edges	5,461	$5{,}236 \pm 95$	$5{,}273 \pm 43$	$5{,}549 \pm 74$	$5{,}288 \pm 25$
F1	—	0.665 ± 0.009	0.656 ± 0.004	0.739 ± 0.010	0.677 ± 0.002
Reciprocity	0.617	0.464 ± 0.069	0.433 ± 0.006	0.555 ± 0.013	0.440 ± 0.006
Assortativity	-0.125	-0.074 ± 0.044	-0.031 ± 0.003	-0.117 ± 0.037	-0.035 ± 0.001
Gini	0.647	0.647 ± 0.000	0.647 ± 0.000	0.647 ± 0.000	0.647 ± 0.000
Clustering	0.407	0.429 ± 0.003	0.400 ± 0.002	0.412 ± 0.004	0.417 ± 0.002
Betw. ρ	—	0.682 ± 0.036	0.670 ± 0.033	0.735 ± 0.045	0.642 ± 0.041
PR ρ	—	0.980 ± 0.002	0.983 ± 0.002	0.987 ± 0.002	0.984 ± 0.001
Eig. ρ	—	0.996 ± 0.000	0.996 ± 0.000	0.996 ± 0.001	0.996 ± 0.000

The GNN's stochastic advantage over the baselines is metric-dependent. On reciprocity, the GNN mean (0.464) exceeds the fitness mean (0.433) and minimum density (0.440), but the GNN's high variance means the distributions overlap substantially. On assortativity, the GNN mean (-0.074) is closer to the ground truth (-0.125) than fitness (-0.031), with the 90% CI [-0.125, 0.012] just reaching the ground truth value. On betweenness centrality, the GNN mean (0.682) is essentially indistinguishable from fitness (0.670), with widely overlapping confidence intervals.

The stochastic maximum entropy method achieves the highest F1 (0.739) and the closest assortativity to ground truth (-0.117), a result that differs markedly from the deterministic maximum entropy in Chapter 5 (which produces 12,323 edges and near-zero assortativity). This occurs because the stochastic version draws approximately 5,549 edges per sample rather than including all possible edges, producing a network with realistic density. This finding underscores that the baseline comparison in Chapter 5 (deterministic, matched to the literature implementations) and the Monte Carlo comparison here (stochastic, matched for fair variance estimation) answer different questions: the former evaluates the methods as typically deployed, while the latter isolates the effect of random variation.

7.1.3 Statistical Tests

For each metric, we test whether the GNN distribution differs significantly from each baseline using Mann-Whitney U tests on the paired 30-seed distributions. On reciprocity, the GNN significantly exceeds fitness ($p < 0.05$) but does not significantly differ from minimum density. On betweenness ρ, no pairwise comparison reaches significance due to the high variance of both distributions. On PageRank ρ, fitness marginally exceeds the GNN (0.983 versus 0.980, $p < 0.05$). The overall pattern is that the GNN's advantages are concentrated in structural metrics (reciprocity, assortativity) while the baselines retain advantages on centrality measures that are dominated by the strength distribution.

7.2 Edge Budget Sensitivity

The GNN+IPF pipeline requires a target edge count to select the top-k most probable edges. In the main analysis, this is estimated from the training-period density. To assess sensitivity to this choice, we vary the edge budget from $0.5\times$ to $1.5\times$ the baseline estimate, holding the

trained model fixed.

Table 7.2: Edge budget sensitivity. Gini is invariant to the budget because IPF determines volume concentration from the marginals.

Multiplier	Edges	F1	Reciprocity	Gini
0.5×	2,824	0.511	0.243	0.647
0.7×	3,829	0.592	0.348	0.647
1.0×	5,250	0.673	0.497	0.647
1.3×	6,601	0.709	0.607	0.647
1.5×	7,501	0.715	0.673	0.647

F1 improves monotonically with the edge budget, reflecting the precision-recall trade-off: more edges increase recall (more true edges are captured) while precision decreases slowly. Reciprocity increases approximately linearly, approaching the ground truth value (0.617) at 1.3–1.5× the baseline budget. The Gini coefficient is invariant at 0.647 across all budget levels, confirming that IPF determines volume concentration independently of the number of edges.

A multi-seed probe (5 seeds) confirms that the optimal multiplier is stable at 1.5× across all seeds tested ($\sigma = 0$). However, the 1.0× baseline represents a conservative choice that avoids overfitting the edge count to the ground truth density, which is an oracle quantity unavailable in practice.

7.3 Temporal Buffer Experiment: e-MID 2006–2007

The policy analysis in Chapter 6 uses a temporal design where buffers are scored on the 2023 network and tested on the 2024 network. To validate this approach on data where the ground truth is known, we apply the same temporal design to the e-MID data, scoring buffers on the 2006 network and testing on the 2007 network.

7.3.1 Ground Truth Arms

Using the true e-MID transaction matrices W_{2006} and W_{2007}, we compute: (i) the in-sample arm, where buffers are scored and tested on W_{2007}, and (ii) the out-of-sample arm, where buffers are scored on W_{2006} and tested on W_{2007}.

The ground truth baseline SR is 100.08 (in the e-MID volume units). The risk reductions and optimism gaps are:

Scoring method	In-sample RR	Out-of-sample RR	Gap
EBA	2.04%	2.00%	0.03 pp
DR-Impact	1.67%	1.60%	0.07 pp
DR-Vuln	2.21%	2.21%	0.00 pp

The optimism gap is negligible for all three methods (≤ 0.07 percentage points), confirming that buffer assignments are temporally stable even on true bilateral networks.

Table 7.3: Temporal buffer experiment on e-MID 2006→2007, 30 Monte Carlo seeds. Optimism gap = in-sample RR − out-of-sample RR.

Method	In-sample RR	Out-of-sample RR	Gap	Jaccard
EBA	$1.35 \pm 0.96\%$	$1.30 \pm 0.91\%$	0.05 ± 0.06 pp	0.979
DR-Impact	$1.00 \pm 0.75\%$	$0.99 \pm 0.72\%$	0.01 ± 0.04 pp	0.999
DR-Vuln	$1.44 \pm 0.96\%$	$1.43 \pm 0.94\%$	0.01 ± 0.11 pp	0.976

7.3.2 Monte Carlo Arms

Repeating the experiment across 30 GNN reconstructions:

The optimism gap is below 0.1 percentage points for all methods and seeds. The deleveraging Jaccard similarity exceeds 0.97, indicating that the same banks are constrained under both in-sample and out-of-sample scoring. The absolute risk reductions are lower in the reconstructed networks (1.0–1.4%) than in the ground truth (1.6–2.2%), reflecting the structural approximation inherent in reconstruction, but the *gap* between in-sample and out-of-sample is similarly negligible. The ORBIS temporal validation (Chapter 6, Table 6.6) confirms this pattern on the 2024 banking system, where the gap is below 0.05 pp with Jaccard = 1.000.

7.4 Distributional Shift: e-MID Training versus ORBIS Inference

The GNN is trained on e-MID transaction data from 2000–2005 and applied to ORBIS balance sheet data from 2024. The 20-year gap and the shift from flow (transaction volume) to stock (balance sheet position) measures raise the question of whether the model's learned representations transfer across this domain change.

7.4.1 Bank-Activity Features

The critical features for edge prediction are [0] and [1]: the log-transformed interbank lending and borrowing volumes. These features drive the gravity scaffold and dominate the GNN's edge ranking.

Table 7.4: Distributional shift on bank-activity features. KS = Kolmogorov-Smirnov statistic. Extrapolation measures the fraction of ORBIS values outside the e-MID training support.

Feature	e-MID mean	ORBIS mean	KS	p	Extrapolation
[0] log1p(lending)	4.865	4.534	0.445	<0.001	0.0%
[1] log1p(borrowing)	3.875	4.820	0.500	<0.001	0.0%

Despite statistically significant distributional differences (KS > 0.44, $p < 0.001$), all ORBIS 2024 values for features [0] and [1] lie within the convex hull of the e-MID training distribution — the model *interpolates* on the dimensions that drive edge prediction. The ORBIS lending feature mean (4.534) is slightly below the e-MID mean (4.865), while the borrowing feature mean (4.820) is slightly above (3.875), reflecting the shift from flow to stock measurement. The log transformation substantially compresses the distributional gap.

7.4.2 Macroeconomic Features

Four of the eight macroeconomic features lie entirely outside the training support: M3 growth (training range 3.7–9.0, ORBIS value 1.5), HICP inflation (1.5–3.0 versus 0.8), Italian unemployment (7.5–10.7 versus 6.6), and the BTP-Bund spread (0.14–0.46 versus 1.32). These reflect the fundamentally different macroeconomic environment of 2024 compared to 2000–2005.

The extrapolation on macro features is mitigated by two factors. First, macro features are common to all nodes on a given day: they shift the probability level uniformly across all node pairs without affecting relative edge rankings. Second, the gravity scaffold — which provides the bulk of the predictive signal — depends only on features [0] and [1], where the model interpolates. Nevertheless, the extrapolation means that the GNN's absolute probability outputs may be miscalibrated, even though the edge ranking is preserved.

7.4.3 Temporal Stability Within e-MID

To contextualize the e-MID-to-ORBIS shift, we compute KS distances between consecutive years within the e-MID training set. For the lending feature, the year-to-year KS ranges from 0.067 to 0.120, while the e-MID-to-ORBIS KS is 0.764 — approximately 6–10$\times$ the natural annual drift. For borrowing, the pattern is similar (annual KS 0.055–0.106, cross-domain 0.484). The cross-domain shift is real and substantial, but it stays within the training envelope (0% extrapolation), meaning the model is asked to interpolate across a larger-than-usual but not unprecedented range of the feature space.

7.5 LGD Sensitivity

The DebtRank algorithm uses a loss-given-default parameter that determines the fraction of exposure lost when a counterparty defaults. The baseline analysis uses LGD = 0.45, the Basel III Foundation IRB default. Table 7.5 reports the sensitivity of the policy conclusions to this assumption.

Table 7.5: LGD sensitivity of systemic risk estimates. All scenarios use the GNN+IPF reconstructed 2024 network.

LGD	No Regulation SR	DR-Vuln SR	Risk reduction
0.30	0.304	0.277	8.7%
0.45	0.546	0.482	11.7%
0.60	0.935	0.765	18.2%

Aggregate systemic risk scales approximately linearly with LGD, as expected from the linear structure of the leverage matrix (Equation 4.14). The qualitative finding that score-based buffers reduce systemic risk is robust: the risk reduction ranges from 8.7% at LGD = 0.30 to 18.2% at LGD = 0.60, with the baseline estimate of 11.7% at LGD = 0.45 sitting in the middle of this range. The increasing effectiveness at higher LGD reflects the amplified contagion dynamics: when losses per default are larger, the deleveraging of systemically important banks prevents proportionally more cascade transmission.

7.6 Feature-Mapping Robustness

The e-MID-to-ORBIS feature bridge maps annual transaction volumes (a flow measure) to balance sheet interbank positions (a stock measure).we test four mapping specifications for features [0]–[3]:

- **Baseline (absolute):** Features [0]–[1] are $\log(1 + \text{IB assets})$ and $\log(1 + \text{IB liabilities})$; feature [2] is $\log(1 + \text{IB assets} + \text{IB liabilities})$ (total interbank volume); feature [3] is the normalized size rank by total assets.
- **Relative exposure:** Features [0]–[1] are $\log(1 + \text{IB assets}/\text{total assets})$ and the liability analogue, normalizing by bank size. Feature [2] is the log of the relative total interbank position; feature [3] remains the normalized size rank.
- **Rank-based:** Features [0]–[1] are log-transformed percentile ranks of IB assets and liabilities within the ORBIS sample, scaled to the training range. Features [2]–[3] duplicate [0]–[1], replacing the total-volume and size-rank signals with the rank-based values. This eliminates the distributional shift between flow and stock magnitudes entirely but also removes the distinct information content of features [2]–[3].
- **Hybrid:** Features [0]–[1] use the baseline absolute mapping; features [2]–[3] use the relative exposure mapping ($\log(1 + \text{IB assets}/\text{total assets})$ and the liability analogue), giving the GNN both absolute size and relative intensity information in place of the total-volume and size-rank features used in the baseline.

For each mapping, we reconstruct the 2024 network using the representative seed (1008), apply the same 2023-scored DR-Vuln buffers, and run DebtRank. The relative mapping is additionally evaluated across all 30 Monte Carlo seeds.

Table 7.6: Feature-mapping robustness: four mapping specifications. All reconstructions use 723 edges and identical marginals. Risk reduction is for the DR-Vuln buffer scenario. Top-10 overlap measures agreement on the ten most systemically important banks.

Mapping	No Reg. SR	DR-Vuln RR	Top-10 overlap
Baseline (absolute)	0.554	12.3%	—
Relative exposure	0.549 ± 0.030	11.9%	80%
Rank-based	0.591	14.4%	80%
Hybrid	0.547	12.0%	90%

All four mappings produce qualitatively identical conclusions. Score-based buffers reduce systemic risk by 11.9–14.4% across all specifications, and the top-10 systemic bank overlap with the baseline exceeds 80% for every alternative. The hybrid mapping achieves the highest overlap (90%) and an SR closest to the baseline (0.547 versus 0.554), suggesting that using both absolute and relative information in features [2]–[3] — rather than duplicating [0]–[1] — provides a more informative signal to the GNN.

The rank-based mapping produces the highest SR (0.591) and the largest risk reduction (14.4%), indicating that the rank transformation shifts the reconstructed topology toward a configuration with somewhat more contagion potential. This is consistent with the rank mapping compressing the heavy-tailed bank size distribution, reducing the dominance of the largest lenders and distributing edges more evenly — a topology that transmits more contagion per default. Despite this level shift, the policy conclusion (buffers reduce risk by approximately 12–14%) and the systemic ranking (80% top-10 agreement) are preserved.

7.7 Sub-Period Training Sensitivity

The GNN is trained on the full 2000–2005 e-MID period. To assess whether the learned structural patterns are specific to this window or reflect stable features of the Italian interbank market, we retrain 30 models on each of four training windows: 2000–2002 (3 years), 2000–2003 (4 years), 2002–2005 (4 years, later period), and 2000–2005 (6 years, baseline). All models are evaluated on the same 2006–2007 ground truth.

Table 7.7: Sub-period training sensitivity: GNN+IPF reconstruction metrics across four training windows (mean $\pm$ sd over 30 seeds). Fitness (stochastic) shown for reference.

Window	F1	Recip.	Assort.	Betw. ρ	PR ρ
2000–2002	$.650 \pm .013$	$.410 \pm .045$	$-.109 \pm .025$	$.479 \pm .075$	$.980 \pm .002$
2000–2003	$.661 \pm .007$	$.441 \pm .036$	$-.089 \pm .025$	$.476 \pm .044$	$.982 \pm .001$
2002–2005	$.672 \pm .007$	$.470 \pm .039$	$-.086 \pm .026$	$.505 \pm .041$	$.980 \pm .001$
2000–2005	$.666 \pm .008$	$.463 \pm .049$	$-.072 \pm .038$	$.499 \pm .061$	$.980 \pm .001$
Fitness	$.628 \pm .010$	$.454 \pm .014$	$-.029 \pm .002$	$.661 \pm .043$	$.978 \pm .004$

Three patterns emerge. First, all four training windows produce reconstruction metrics in the same range: F1 spans 0.650–0.672, reciprocity 0.410–0.470, and PageRank ρ 0.980–0.982. No window produces a catastrophic failure, confirming that the structural patterns learned by the GNN — core-periphery topology, degree heterogeneity, reciprocity — are stable features of the Italian interbank market rather than artifacts of a specific period.

Second, reconstruction quality improves modestly with both the *amount* and *recency* of training data. The 2002–2005 window (4 years, most recent) achieves the highest F1 (0.672) and reciprocity (0.470), outperforming the 2000–2005 baseline (6 years) on both metrics. This suggests that temporal proximity to the test period matters more than sample size, consistent with the gradual evolution of interbank market structure documented in the monthly evaluation (Chapter 5).

Third, the Gini coefficient and eigenvector centrality correlation are invariant to the training window (0.647 ± 0.000 and 0.996 ± 0.000 respectively), confirming that these properties are determined by IPF and the strength distribution rather than the GNN's learned topology. Figure 7.1 visualizes the full comparison.

7.8 Limitations of the Robustness Analysis

Two robustness dimensions remain untested. First, all experiments use the same GNN architecture (two-layer TransformerConv); robustness to architectural choices (GraphSAGE, VGAE, heterogeneous graph transformers) remains untested. Second, the stochastic baseline comparison reveals that probabilistic maximum entropy achieves higher F1 and closer assortativity to ground truth than the GNN, a result that merits further investigation in future work but falls outside the scope of the deterministic baseline comparison that defines the main contribution.

Figure 7.1: Sub-period training sensitivity. Each panel shows a reconstruction metric across four training windows (blue, with ± 1 sd error bars) against the fitness baseline (red band) and ground truth (green dotted line where applicable). The GNN produces stable outputs across all windows, with modest improvements from more recent training data.

Chapter 8

Conclusion

This thesis has developed and evaluated a hybrid pipeline for interbank network reconstruction that combines a Graph Neural Network trained on historical transaction data with Iterative Proportional Fitting to produce weighted directed networks consistent with observed balance sheet marginals. The pipeline has been validated against three established baselines on the Italian interbank market and applied to a temporal analysis of macroprudential capital buffers on the 2024 Italian banking system, with buffers scored on the 2023 network and tested out-of-sample on 2024. We summarize the main contributions, discuss the limitations, and outline directions for future research.

8.1 Summary of Contributions

The thesis makes four contributions.

The first is **methodological**. The GNN+IPF pipeline addresses a structural gap in the network reconstruction literature: classical methods reconstruct each period's network from scratch using only contemporaneous marginals, discarding the topological regularities that persist across time in real interbank markets. By training a GNN on historical e-MID transaction data, we extract these regularities and embed them as informed priors for the IPF algorithm. The two-stage design exploits the comparative advantage of each component: the GNN learns *where* money flows, while IPF ensures *how much* matches the observed balance sheet constraints. The gravity scaffold provides the GNN with a structurally plausible input topology without exposing it to the edges it is tasked with predicting, maintaining a clean separation between input and label throughout training and inference. A loss-weight sensitivity analysis over 15 configurations identifies the operating point that balances edge classification accuracy against structural fidelity.

An ablation study confirms that the GNN encoder contributes learned structural patterns beyond what the gravity scaffold provides. The strongest evidence comes from metrics entirely absent from the training objective: the Gini coefficient of out-strength is matched exactly (error reduction from 0.088 to 0.000 over gravity-only), PageRank preservation improves from $\rho = 0.891$ to 0.983 (+84%), and eigenvector centrality from 0.907 to 0.996 (+96%). Because these metrics are never presented to the model during training, the improvements rule out circularity. The one untrained structural metric where the scaffold outperforms the GNN is clustering (error 0.012 versus 0.023), attributable to the structural losses reshaping the edge distribution in ways that inflate the triangle count.

The second contribution is **empirical**. On the out-of-sample reconstruction of the 2006–2007 Italian interbank network, GNN+IPF achieves an exact match on the Gini coefficient

of out-strength while the fitness model deviates by 0.023, and significantly outperforms all baselines on centrality preservation (Wilcoxon $p = 0.004$ on eight non-mechanical centrality measures). The mean Spearman rank correlation is 0.926 for GNN+IPF versus 0.905 for fitness, with the advantage concentrated in authority score (+0.071), betweenness (+0.046), and PageRank (+0.012). The fitness model retains a modest edge on F1 score (0.692 versus 0.674), reflecting its direct calibration to marginals. On reciprocity and assortativity, both methods perform comparably at the bi-annual aggregation level, with the GNN providing a modest structural advantage that is persistent across monthly sub-periods (winning 15/24 months on reciprocity, 14/24 on assortativity). The monthly evaluation reveals a temporal aggregation effect — both methods overestimate monthly reciprocity while underestimating bi-annual reciprocity — and documents the onset of the 2007 financial crisis in the network topology, where the GNN's learned priors from the stable 2000–2005 period produce larger errors during the structural break than the agnostic fitness model.

The third contribution is the **policy application**. Using a temporal design where buffers are scored on the 2023 reconstructed network and tested out-of-sample on the 2024 network, we show that score-based macroprudential capital buffers reduce aggregate systemic risk by approximately 12% relative to the unregulated baseline, achieved through the deleveraging of 6–9 marginally capitalized banks. The three scoring mechanisms evaluated (EBA indicator-based, DebtRank impact, DebtRank vulnerability) produce similar outcomes (11.9–12.7% risk reduction), confirming that the well-capitalized nature of the Italian banking system limits the differentiation between scoring methods. The temporal validation confirms that in-sample buffer calibration does not overstate effectiveness: the optimism gap between in-sample and out-of-sample scoring is below 0.05 percentage points across all methods, and the set of deleveraging banks is identical across years (Jaccard = 1.000).

A notable finding is that the GNN+IPF and fitness reconstructions produce nearly identical systemic risk estimates on the 2024 network (SR = 0.554 versus 0.552, a gap of 0.4%). This convergence occurs because both methods share the same marginals (enforced by IPF) and the same edge count, and the contagion dynamics in this well-capitalized system are dominated by exposure volume rather than exposure pattern. The structural advantages documented in Chapter 5 — higher centrality preservation, exact Gini match, more realistic degree distribution — improve the fidelity of the reconstructed network but do not change the direction or magnitude of the policy recommendations. Whether this convergence holds in a stressed system with binding capital constraints and larger shocks is an open question.

The fourth contribution is the **robustness analysis**. Seven robustness tests confirm the stability of the pipeline's outputs: the 30-seed Monte Carlo shows that reconstruction metrics are stable across random initializations (Gini sd $< 10^{-7}$, centrality correlations sd < 0.04); the edge-budget sweep confirms that the Gini coefficient is invariant to the selection threshold; the temporal buffer experiment validates the out-of-sample design on e-MID ground truth data (optimism gap < 0.1 pp); the distributional shift analysis confirms that the GNN interpolates on the bank-activity features that drive edge prediction (0% extrapolation); the LGD sensitivity and feature-mapping robustness checks confirm that all qualitative policy conclusions survive perturbation (risk reduction 8.7–18.2% across the LGD range, 80–90% top-10 overlap across four alternative feature specifications); and the sub-period training analysis shows that reconstruction quality is stable across four training windows spanning 2000–2005 (F1 range 0.650–0.672, PageRank ρ range 0.980–0.982), with modest improvements from more recent training data.

8.2 Limitations

The thesis is subject to several limitations.

Static reconstruction. The pipeline produces a single network snapshot from cross-sectional balance sheet data. It does not model the dynamic evolution of interbank relationships, the endogenous rewiring of lending in response to capital requirements, or the second-round effects of deleveraging on other banks' funding. The counterfactual analysis captures the structural effect of capital buffers but not the behavioral feedback effects that an agent-based model like that of Gurgone and Iori (2022) would capture. This limitation is particularly relevant for the comparison of scoring mechanisms: in a dynamic setting, DebtRank-based scoring might outperform EBA scoring more decisively because it targets the banks whose deleveraging has the largest knock-on effects, but our static framework cannot detect this advantage.

Feature bridge and distributional shift. The application of a GNN trained on e-MID transaction flows (2000–2005) to ORBIS balance sheet stocks (2024) involves a 20-year domain gap. The distributional shift analysis (Chapter 7) confirms that the model interpolates on the critical bank-activity features but extrapolates on four macroeconomic features. The feature-mapping robustness check shows that the qualitative conclusions survive an alternative specification, but we cannot verify that the GNN's learned representations transfer with full fidelity across this gap. The structural patterns of the Italian interbank market have evolved since 2005, and some patterns learned from pre-crisis data may no longer hold. The monthly evaluation's documentation of the 2007 H2 structural break reinforces this concern: the GNN's priors produce larger errors during regime changes than the structurally agnostic fitness model.

Small-shock regime. The policy analysis operates in what Acemoglu et al. (2015) characterize as the small-shock regime, where individual bank defaults propagate without triggering system-wide collapse. The convergence between GNN+IPF and fitness systemic risk estimates (0.4% gap) may not hold in the large-shock regime, where network topology becomes the dominant determinant of contagion dynamics. The 2024 Italian banking system, with mean CET1 above 20%, is firmly in the small-shock regime for the scenarios considered.

RWA approximation. Risk-weighted assets are estimated as 50% of total assets for all banks, ignoring heterogeneity in asset composition. This simplification affects which banks are forced to deleverage. Bank-level Pillar 3 disclosures would permit a more accurate calibration.

EBA-impact correlation. The correlation between EBA scores and DebtRank impact scores (0.885) remains high, suggesting that IPF-based reconstruction mechanically couples size with network importance. This limits the ability of DebtRank-based scoring to identify systemically important institutions that are not also among the largest.

Single market. The pipeline is trained and validated on a single national interbank market. The structural regularities learned from e-MID may be specific to the Italian institutional context and may not transfer to markets with different structures.

Stochastic baseline comparison. The Monte Carlo analysis in Chapter 7 reveals that stochastic maximum entropy — which draws a realistic number of edges rather than including all possible links — achieves higher F1 and closer assortativity to ground truth than the GNN. This finding, which differs from the deterministic comparison in Chapter 5, suggests that the baseline landscape is richer than the standard deterministic implementations imply and that the GNN's advantages may be narrower when baselines are given the benefit of stochastic sampling.

8.3 Future Research

Several extensions could address the limitations identified above.

Temporal GNN architectures. Replacing the static TransformerConv encoder with a temporal GNN that processes sequences of daily graphs would allow the model to capture the evolution of interbank relationships over time. This could improve reconstruction accuracy during periods of structural change and enable time-varying edge probabilities rather than a single aggregate. The monthly evaluation's documentation of the 2007 crisis onset provides a natural test case for temporal models.

Stochastic baseline investigation. The Monte Carlo finding that probabilistic maximum entropy outperforms the GNN on several metrics warrants systematic investigation. A fair comparison of the GNN against optimally calibrated stochastic baselines — rather than the deterministic implementations standard in the literature — would clarify the GNN's marginal contribution and potentially motivate hybrid approaches that combine learned structural priors with stochastic sampling.

Cross-border extension. Extending the pipeline to multi-country networks using TARGET2 payment data or multi-country e-MID data would permit the analysis of cross-border contagion channels. The GNN architecture naturally accommodates heterogeneous node sets, and the feature bridge could incorporate country-level macro variables.

Dynamic ABM integration. The static DebtRank analysis could be complemented by an agent-based model in which banks endogenously adjust interbank lending in response to capital requirements. The GNN+IPF reconstruction would provide the initial network topology, and the ABM would simulate the dynamic evolution under different policy scenarios, capturing the behavioral feedback effects that the static framework misses.

Large-shock regime. The convergence between GNN+IPF and fitness systemic risk estimates in the small-shock regime raises the question of whether the topology effect re-emerges under larger shocks. Simulating multi-bank default scenarios, correlated shocks, or fire-sale dynamics on the reconstructed network would test whether the GNN's structural advantages translate into different contagion outcomes when the system is pushed beyond the well-capitalized baseline.

Validation with supervisory data. The ultimate test of any reconstruction method is comparison against the true bilateral exposure matrix, available to supervisory authorities through large exposure reporting (FinRep/CoRep) and credit registers. A collaboration with the Banca d'Italia or the ECB's Single Supervisory Mechanism that provided access to the true 2024 bilateral matrix would allow direct validation and a definitive comparison between methods.

Bibliography

Acemoglu, D., Ozdaglar, A., and Tahbaz-Salehi, A. (2015). Systemic risk and stability in financial networks. *American Economic Review*, 105(2):564–608.

Allen, F. and Gale, D. (2000). Financial contagion. *Journal of Political Economy*, 108(1):1–33.

Anand, K., Craig, B., and von Peter, G. (2015). Filling in the blanks: Network structure and interbank contagion. *Quantitative Finance*, 15(4):625–636.

Anand, K., van Lelyveld, I., Banai, Á., Friedrich, S., Garratt, R., Hałaj, G., Fique, J., Hansen, I., Martínez Jaramillo, S., Lee, H., Molina-Borboa, J. L., Nobili, S., Rajan, S., Salakhova, D., Silva, T. C., Silvestri, L., and Stancato de Souza, S. R. (2018). The missing links: A global study on uncovering financial network structures from partial data. *Journal of Financial Economics*, 130(2):367–390.

Bacharach, M. (1965). Estimating nonnegative matrices from marginal data. *International Economic Review*, 6(3):294–310.

Banca d'Italia (2023). Banche e istituzioni finanziarie: Articolazione territoriale — metodi e fonti, note metodologiche. Statistical methodology note.

Bardoscia, M., Battiston, S., Caccioli, F., and Caldarelli, G. (2015). Debtrank: A microscopic foundation for shock propagation. *PLoS ONE*, 10(6):e0130406.

Battiston, S., Puliga, M., Kaushik, R., Tasca, P., and Caldarelli, G. (2012). DebtRank: Too central to fail? Financial networks, the FED and systemic risk. *Scientific Reports*, 2(1):541.

Brandes, U. (2001). A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25(2):163–177.

Cimini, G., Squartini, T., Musmeci, N., Puliga, M., Gabrielli, A., Garlaschelli, D., Battiston, S., and Caldarelli, G. (2015). Reconstructing topological properties of complex networks using the fitness model. *Physical Review E*, 92(4):040802.

Cinelli, C. (2017). NetworkRiskMeasures: Risk measures for financial networks. R package version 0.1.4, `https://CRAN.R-project.org/package=NetworkRiskMeasures`.

Cont, R., Moussa, A., and Santos, E. B. e. (2013). Brazilian interbank networks: Assessing systemic risk. *Quantitative Finance*, 13(6):867–885.

Degryse, H. and Nguyen, G. (2007). Interbank exposures: An empirical examination of contagion risk in the Belgian banking system. *International Journal of Central Banking*, 3(2):123–171.

Eisenberg, L. and Noe, T. H. (2001). Systemic risk in financial systems. *Management Science*, 47(2):236–249.

European Banking Authority (2014). Guidelines on the criteria to determine the conditions of application of Article 131(3) of Directive 2013/36/EU (CRD) in relation to the assessment of other systemically important institutions (O-SIIs). Technical Report EBA/GL/2014/10, European Banking Authority.

Fricke, D. and Lux, T. (2015). Core–periphery structure in the overnight money market: Evidence from the e-MID trading platform. *Computational Economics*, 45(3):359–395.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. (2017). Neural message passing for quantum chemistry. *Proceedings of the 34th International Conference on Machine Learning*, pages 1263–1272.

Gurgone, A. and Iori, G. (2022). Macroprudential capital buffers in heterogeneous banking networks: Insights from an ABM with liquidity crises. *European Journal of Finance*, 28(13–15):1399–1445.

Hamilton, W., Ying, Z., and Leskovec, J. (2017). Inductive representation learning on large graphs. *Advances in Neural Information Processing Systems*, 30.

Iori, G., De Masi, G., Precup, O. V., Gabbi, G., and Caldarelli, G. (2008). A network analysis of the Italian overnight money market. *Journal of Economic Dynamics and Control*, 32(1):259–278.

Kipf, T. N. and Welling, M. (2016). Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*.

Kipf, T. N. and Welling, M. (2017). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*. Proceedings of the Fifth International Conference on Learning Representations (ICLR 2017).

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. *Proceedings of the Seventh International Conference on Learning Representations (ICLR 2019)*.

Macchiati, V., Mazzarisi, P., and Garlaschelli, D. (2022). Interbank network reconstruction enforcing density and reciprocity. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(5):053401.

Mistrulli, P. E. (2011). Assessing financial contagion in the interbank market: Maximum entropy versus observed interbank lending patterns. *Journal of Banking & Finance*, 35(5):1114–1127.

Newman, M. E. J. (2002). Assortative mixing in networks. *Physical Review Letters*, 89(20):208701.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. Stanford InfoLab Technical Report.

Petropoulos, A., Siakoulis, V., Lazaris, P., and Chatzis, S. (2020). Re-constructing the interbank links using machine learning techniques: An application to the Greek interbank market. *Journal of Network Theory in Finance*, 6(3):1–32.

Precup, O., Iori, G., and Gabbi, G. (2005). The microstructure of the Italian overnight money market. Working paper, King's College London.

Sheldon, G. and Maurer, M. (1998). Interbank lending and systemic risk: An empirical analysis for switzerland. *Swiss Journal of Economics and Statistics*, 134(IV):685–704.

Shi, Y., Huang, Z., Feng, S., Zhong, H., Wang, W., and Sun, Y. (2021). Masked label prediction: Unified message passing model for semi-supervised classification. *arXiv preprint arXiv:2009.03509*. Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21).

Squartini, T., Caldarelli, G., Cimini, G., Gabrielli, A., and Garlaschelli, D. (2018). Reconstruction methods for networks: The case of economic and financial systems. *Physics Reports*, 757:1–47.

Tinbergen, J. (1962). Shaping the world economy: Suggestions for an international economic policy.

Upper, C. (2011). Simulation methods to assess the danger of contagion in interbank markets. *Journal of Financial Stability*, 7(3):111–125.

Upper, C. and Worms, A. (2004). Estimating bilateral exposures in the German interbank market: Is there a danger of contagion? *European Economic Review*, 48(4):827–849.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. (2018). Graph attention networks. *arXiv preprint arXiv:1710.10903*. Proceedings of the Sixth International Conference on Learning Representations (ICLR 2018).

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442.

Zhang, M. and Chen, Y. (2018). Link prediction based on graph neural networks. *Advances in Neural Information Processing Systems*, 31.