Impact of Graph Fusion Strategies on Multi-Layer Open-Source Ecosystem Networks

A Comparative Study on Union-based and Consensus-based Approaches

[Jingsen Zhang, Zeqiang Wang, Zixi Chen]

[NYU Shanghai]

[December 28, 2025]

Motivation & Research Context

Core Message: Open-Source Software (OSS) ecosystems are inherently multi-dimensional.

Key Content:

1. **Heterogeneous Signals:** Collaboration (PRs/Issues), Activity (Events), and Popularity (Stars/OpenRank).
2. **The Problem:** Current studies often "flatten" these signals into a single graph prematurely.
3. **Positioning:** Graph fusion is a significant modeling choice, not a trivial preprocessing detail.

Problem Formalization

Core Message: Defining the Node-Aligned Multi-Layer Graph.

Formal Definition:

1. **Input:** A set of graphs $\mathcal{G} = \{G_1, G_2, \ldots, G_k\}$ where all layers share a consistent node set V.
2. **Challenge:** Layers differ in edge semantics and weight distributions.
3. **Objective:** Construct a unified representation G^* that preserves meaningful relational signals.

Why Existing Practices Are Insufficient

Core Message: Moving beyond simple embedding-based similarity.

Critical Limitations:

1. **Embedding Gap:** High semantic similarity in embedding space does not equal actual interaction or co-occurrence in reality.
2. **Methodological Blind Spot:** Most studies default to one fusion strategy without comparing the systematic bias introduced.

Research Questions (RQs)

Objective: Systematically investigating the trade-offs of fusion.

The Three Questions:

1. **RQ1 (Structural Impact):** How do fusion strategies alter network density and degree distributions?
2. **RQ2 (Trade-offs):** What is the balance between maximizing information coverage and filtering noise?
3. **RQ3 (Stability):** To what extent do core nodes and community structures remain invariant?

Data Source & Pipeline

Context: A reproducible setting using OpenDigger (GitHub event-derived data).

1. **Layer Details:** Stars, Forks, OpenRank, and Developer Collaboration.
2. **Implementation:** Shared identifiers in structured JSON files enable straightforward cross-layer node alignment.

Methodology: Comparison of Fusion Strategies

Strategy 1: Union-based Fusion

Formula: $E_{union} = \cup E_i$

Aim: Maximize information coverage and include weak signals.

Strategy 2: Consensus-based Fusion

Formula: $E_{cons} = \cap E_i$

Aim: Multi-view support as a proxy for reliability; reduce noise.

Evaluation Framework

Metric Layers:

1. **Macro:** Degree distribution and clustering coefficients.
2. **Meso:** Stability of community detection results (e.g., Louvain/Infomap).
3. **Micro:** Ranking stability of top nodes (e.g., influential repositories).
4. **Qualitative:** Visual inspection of core subgraphs.

Expected Outcomes & Significance

Deliverables: A reproducible fusion pipeline and comparative report.

Significance: Provides a transparent roadmap for researchers to select fusion strategies based on specific analysis goals (e.g., identifying stable cores vs. emerging projects).

Selected References

Key Literature (Must Include):

- Kivelä et al. (2014). Multilayer Networks. *Journal of Complex Networks*.
- De Domenico et al. (2014). Multilayer Networks: Structure and Function. *Physics Reports*.
- Hamilton (2020). Graph Representation Learning.
- Yang et al. (2025). A Survey on Multi-View Knowledge Graphs. *IJCAI*.

Thank You!

Q & A