Task 1: Dataset Exploration of MetaFam

Graph Properties, Centrality Analysis, and Structural Insights

"Makes a diff'rence, havin' a decent family." — Rubeus Hagrid

Contents

1 Introduction and Methodology

1.1 Objective

The first step in any data science task is understanding the data. This report documents a thorough exploration of the MetaFam Knowledge Graph, answering the fundamental questions — how many people, how many relationship types, what distributions arise — and then going significantly deeper into graph-theoretic properties, data quality verification, and family-specific structural analysis.

★ Exploration beyond Requirements

Beyond the required statistics and visualization, I conducted several additional analyses:

- **Data quality verification** — reciprocity analysis, temporal cycle detection, gender consistency checks, and transitivity validation
- **Family-specific importance metrics** — Progenitor Score, Ancestor Score, and Lineage Centrality as domain-appropriate alternatives to standard centrality measures
- **Structural isomorphism testing** — Weisfeiler–Lehman hashing to determine whether the 50 families share identical structures
- **Triangular motif analysis** — discovering the "DNA" of relationship patterns
- **Marriage and fertility analysis** — detecting couples and analyzing children-per-couple distributions
- **Interactive PyVis visualization** — a full interactive HTML graph of the entire knowledge graph
- **Ancestor classification** — distinguishing true founders from nodes with missing parent data

1.2 Data Loading

The dataset was loaded from `train.txt`, a tab-separated file containing (head, relation, tail) triples. A basic sanity check confirmed **zero duplicate triples** in the dataset.

2 Basic Dataset Statistics

2.1 Overview

Table 1: MetaFam Dataset Summary

Metric	Value
Total Triples	13,821
Unique People (Nodes)	1,316
Unique Relationship Types	28
Duplicate Triples	0
Number of Families (Components)	50
Average Family Size	26.3 ± 0.5
Family Size Range	26–27

The graph is **not connected** — it consists of 50 weakly connected components, each representing an independent family. The remarkably uniform family sizes (26–27 members) immediately suggest this is a synthetically generated dataset with controlled parameters.

2.2 Relationship Distribution

The 28 relationship types span four broad categories: parent–child, sibling, grandparent, and extended family.

Table 2: Top 10 Relationship Types by Frequency

Relation	Count
`grandsonOf`	814
`grandmotherOf`	813
`grandfatherOf`	813
`granddaughterOf`	812
`motherOf`	733
`fatherOf`	733
`sisterOf`	636
`daughterOf`	628
`greatGrandsonOf`	624
`greatGrandmotherOf`	617

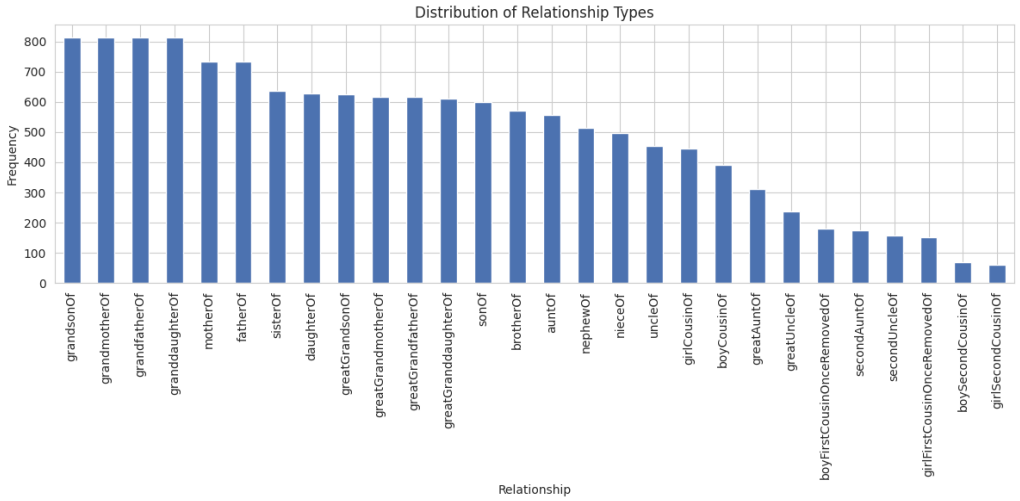

Figure 1: Distribution of all 28 relationship types. Grandparent relations are most frequent, followed by parent–child and sibling relations. Extended family relations (second cousins, cousins once removed) are least frequent.

Insight

Grandparent relations (`grandsonOf`, `grandmotherOf`, etc.) are the most frequent edge type — more common than direct parent–child edges. This is because each grandparent connects to *all* grandchildren across multiple children's families, creating a multiplicative effect. A grandparent with 3 children who each have 2 kids produces 6 grandparent edges but only 3 parent edges.

3 Graph Structure and Properties

3.1 Density and Connectivity

Table 3: Graph-Theoretic Properties

Property	**Value**
Graph Density	0.007987
Weakly Connected	No
Number of Components	50
Diameter (largest family)	3
Radius (largest family)	2
Average Path Length	1.470

The graph is extremely sparse (density < 0.01), which is expected for family trees — each person is related to only a small fraction of all people. Within families, however, the picture is different: average within-family density is 0.415, meaning family members are connected to roughly 42% of their relatives.

3.2 Clustering Coefficient

Key Finding

The family graph is 50.5× more clustered than a random graph:

- Global Clustering Coefficient: **0.7696**
- Average Local Clustering Coefficient: **0.7908**
- Random Graph Clustering (same n, m): 0.0152

This extreme clustering is a signature of family structure: if A is related to B and B is related to C, then A is very likely related to C as well. Families are inherently transitive.

3.3 Path Length Analysis

Within the largest family component (27 people):

- **Average path length: 1.470** — any two family members are on average ∼1.5 relationships apart
- **Diameter: 3** — the most distant relatives are only 3 steps apart
- **Radius: 2** — center nodes can reach everyone within 2 hops

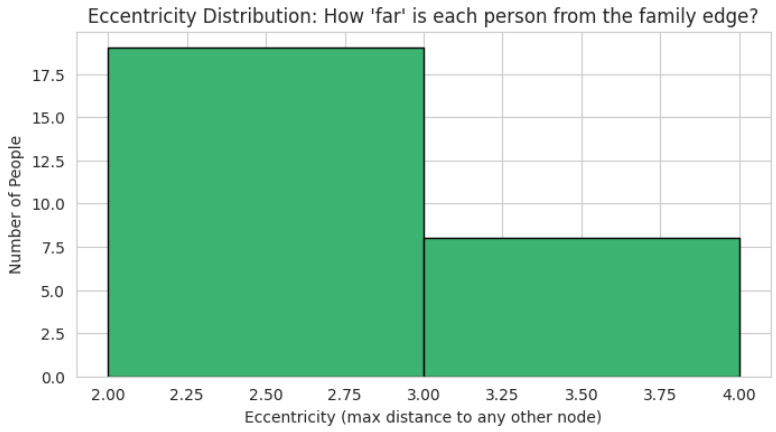

Figure 2: Eccentricity distribution within the largest family. Most nodes have eccentricity 2 (can reach everyone in 2 hops), with a few peripheral nodes at eccentricity 3.

Insight

The short diameter (3) and low average path length (1.47) reflect the dense interconnectedness within families. In a knowledge graph where sibling, cousin, uncle, and grandparent edges all exist explicitly, most family members are directly connected — you rarely need to traverse intermediate nodes.

3.4 Degree Distribution

Table 4: Degree Statistics

Statistic	Value
Minimum Degree	1
Maximum Degree	45
Mean Degree	21.00
Median Degree	22

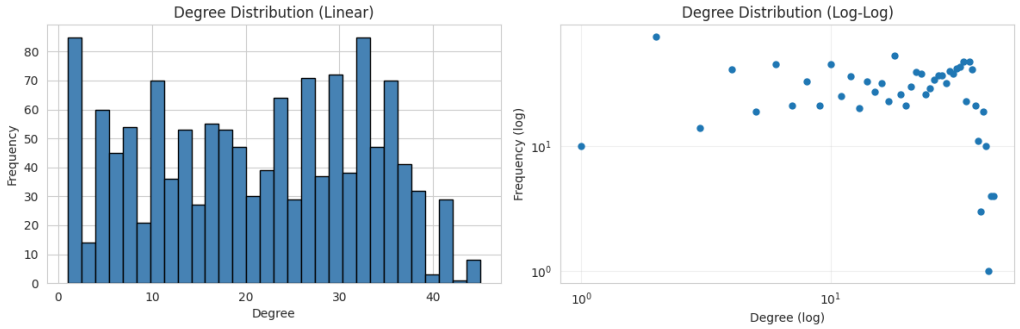

Figure 3: Degree distribution in linear (left) and log-log (right) scales. The distribution is unimodal and concentrated, *not* following a power law. This confirms the graph is not scale-free — consistent with a synthetic family graph where degree is determined by family position, not preferential attachment.

The mean ≈ median ≈ 21 indicates a symmetric, non-skewed distribution. The log-log plot shows no linear trend, confirming this is **not a scale-free network**. This makes biological sense: degree in a family graph is determined by one's structural position (number of siblings, children, cousins), not by any "rich-get-richer" dynamic.

3.5 Directed Graph: In-Degree vs. Out-Degree

Table 5: Directed Degree Statistics

	Value	Top Node
Mean In-Degree	10.50	dominik1036 (23)
Mean Out-Degree	10.50	oskar133 (22)
Max Asymmetry (outward)	+7	sophie309, konstantin493
Max Asymmetry (inward)	−5	natalie811, simon1197

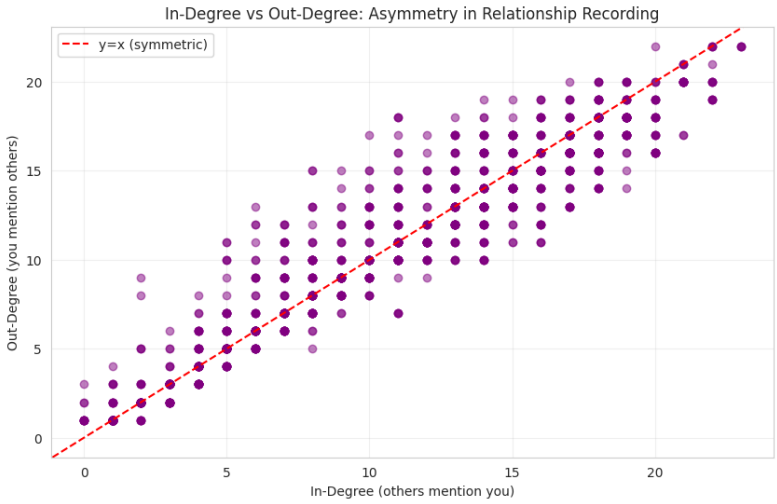

Figure 4: In-degree vs. out-degree scatter plot. Most nodes cluster near the $y = x$ line (symmetric relationship recording). Nodes above the line mention others more than they are mentioned; nodes below are mentioned more.

3.6 Assortativity

The degree assortativity coefficient is **0.1213** (weakly positive), indicating a mild tendency for high-degree nodes to connect to other high-degree nodes. In family terms: people with many recorded relationships tend to be related to others who also have many relationships — which makes sense, as large families produce high degree for all their members.

4 Centrality and Important Nodes

4.1 Standard Centrality: Degree

Table 6: Top 5 Most Connected People (by Total Degree)

Node	Degree
dominik1036	45
magdalena1044	45
oliver1045	45
lisa1035	45
oskar133	44

Standard degree centrality identifies people with the most explicit relationship edges. However, in a family KG where relationship types vary in semantic importance, raw degree conflates a `fatherOf` edge with a `secondCousinOf` edge.

4.2 Family-Specific Importance Metrics

★ *Custom metrics designed for genealogical importance.*

I defined three domain-appropriate centrality measures using the parent→child directed acyclic graph:

1. **Progenitor Score**: Number of descendants (how many people descend from you)
2. **Ancestor Score**: Number of ancestors (how deep is your recorded lineage)

3. **Lineage Centrality**: Descendants + Ancestors (who spans the most generations)

Table 7: Top Progenitors (Most Descendants)

Node	Descendants	Generation
emma7	17	0
moritz8	17	0
marie113	17	0
daniel114	17	0
sofia191	17	0

Table 8: Deepest Lineage (Most Ancestors)

Node	Ancestors	Generation
elena257	20	6
valerie260	20	6
sarah251	20	6
hannah39	18	5
dominik44	18	5

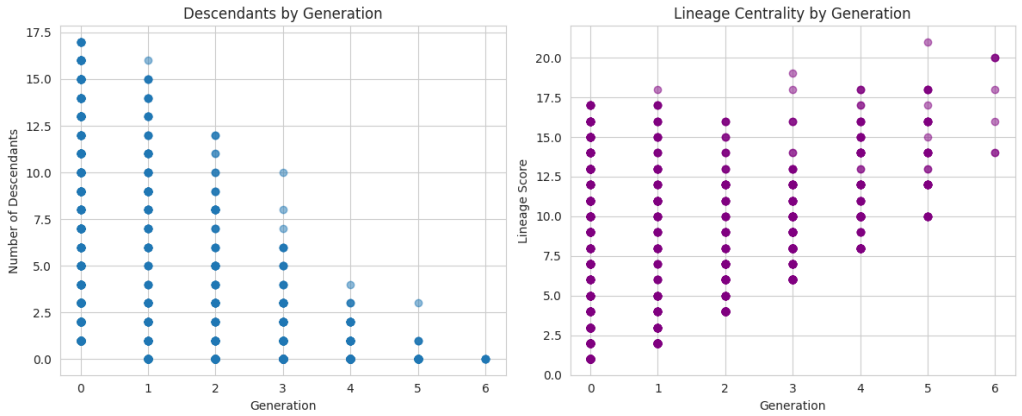

Figure 5: Left: Descendants by generation — founders (Gen 0) have the most descendants, decaying monotonically. Right: Lineage Centrality peaks at middle generations, which span the most total lineage.

Insight

Progenitor Score peaks at Generation 0 (founders have the most descendants), while **Lineage Centrality peaks at middle generations** (Gen 2–3) who have both significant ancestry above and descendants below. This mirrors real genealogy: the "most important" family members depend on your definition — founders who started the lineage, or middle-generation members who connect the most people.

5 Generational Analysis

5.1 Generation Detection

I built a directed "time graph" containing only parent→child edges (`fatherOf`, `motherOf`, and the inverses `sonOf`, `daughterOf` flipped). Generations were assigned via topological sort: roots (in-degree 0) get Generation 0, and each child is assigned one generation beyond its highest-generation parent.

Table 9: Population by Generation

Generation	People
Gen 0 (Founders)	495
Gen 1	215
Gen 2	192
Gen 3	206
Gen 4	146
Gen 5	55
Gen 6	7
Total	**1,316**

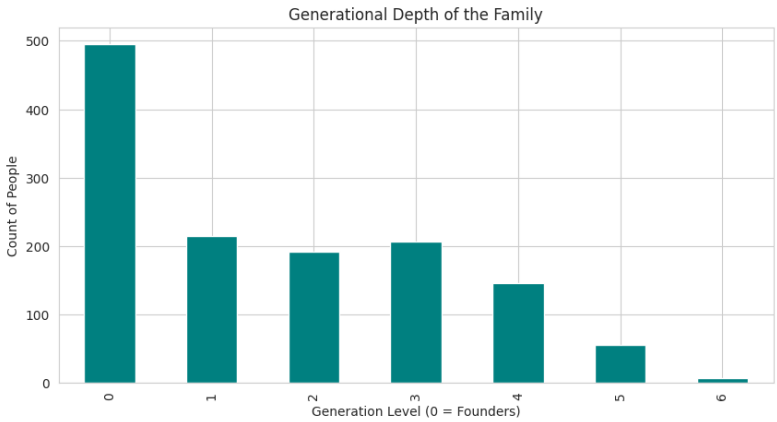

Figure 6: Population by generation. Gen 0 dominates with 495 members, tapering to just 7 at Gen 6. The large Gen 0 count prompted further investigation.

5.2 The 495 Ancestors: True Founders vs. Missing Data

The 495 people assigned to Generation 0 seemed unexpectedly large. I investigated whether all of them are truly founding ancestors.

Key Finding

Only 114 of 495 "ancestors" are true founders. The remaining 381 (77%) have grandchildren recorded in the data, meaning they *should* have parents — those parent edges simply aren't in the dataset.

- True founding ancestors (no grandchildren): **114** (23.0%)
- Ancestors with missing parent data: **381** (77.0%)

The "ancestors" who hold `grandfatherOf` or `grandmotherOf` relationships clearly occupy middle generations in reality, but appear as Gen 0 because their own parents are unrecorded.

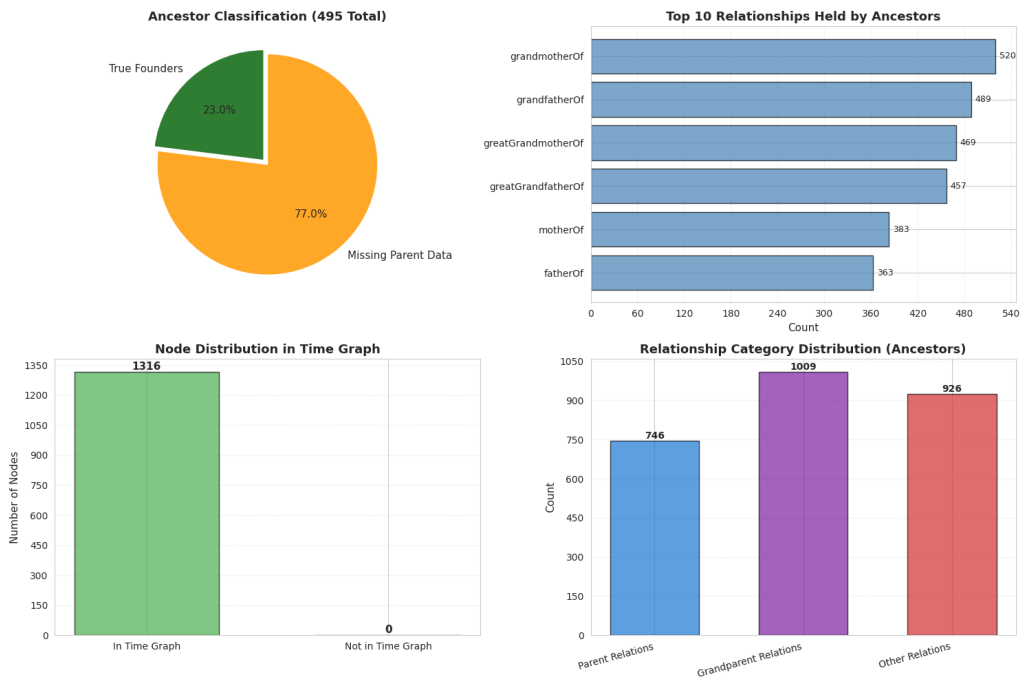

Figure 7: Ancestor classification. Top-left: Only 23% are true founders. Top-right: Many "ancestors" hold grandparent and great-grandparent relations, indicating they are mid-generation with missing parents. Bottom: All 1,316 nodes appear in the time graph, but 381 are misclassified as Gen 0.

6 Data Quality and Consistency

6.1 Gender Inference and Consistency

Gender was inferred from relationship names: relations containing "father," "brother," "son," "uncle," "nephew," "grandfather," "grandson" indicate male heads; "mother," "sister," "daughter," "aunt," "niece," "grandmother," "granddaughter" indicate female heads.

Table 10: Gender Distribution

Gender	Count
Male	646
Female	670
Gender Contradictions	**0**

All 1,316 nodes were successfully classified with **zero contradictions** — no person appears as both male and female across different edges.

6.2 Temporal Consistency: Cycle Detection

I checked whether the parent→child directed graph contains cycles, which would represent a temporal paradox (someone being their own ancestor).

Result: No cycles found. The time graph is a valid DAG (Directed Acyclic Graph).

Note: The full multigraph naturally contains "cycles" from symmetric relations (`sisterOf(A,B)` + `sisterOf(B,A)`), with 2,042 such symmetric edges. These are expected and correct.

6.3 Transitivity Validation

I verified the compositional rule: if A is parent of B and B is parent of C, then A should be grandparent of C.

Key Finding

Grandparent transitivity: 100% complete. Of 1,294 expected grandparent edges (derived from parent chains), all 1,294 exist in the data. The knowledge graph perfectly encodes this logical rule.

6.4 Reciprocity Analysis

For each edge with a known inverse (e.g., `fatherOf` ↔ `sonOf`/`daughterOf`), I checked whether the reciprocal edge exists.

Table 11: Reciprocity Analysis Summary

Metric	Value
Edges checked	9,172
Reciprocal edges found	8,582
Missing reciprocals	590
Overall reciprocity rate	93.57%

Table 12: Reciprocity Rate by Relationship Type (Incomplete Relations Only)

Relation	Total	Missing	Rate
`motherOf`	733	289	60.6%
`fatherOf`	733	125	82.9%
`sonOf`	600	88	85.3%
`daughterOf`	628	88	86.0%
All other relations	—	0	100.0%

Figure 8: Reciprocity analysis. Top-left: Overall 93.6% reciprocity. Top-right: Rate by relation type — only parent–child relations have gaps. Bottom-left: `motherOf` has the most missing reciprocals (289). Bottom-right: Missing count correlates with edge frequency for parent–child types only.

Key Finding

All 590 missing reciprocals are parent–child edges. Every other relationship type (siblings, grandparents, aunts/uncles, cousins) achieves 100% reciprocity. The asymmetry within parent–child relations is striking: `motherOf` has only 60.6% reciprocity (289 missing) while `fatherOf` achieves 82.9% (125 missing).
This systematic incompleteness is not random — it reveals deliberate design. As confirmed in later tasks, these 590 missing edges form precisely the test set for link prediction (Task 4).

7 Family Structure Analysis

7.1 Component Analysis

All 50 families have remarkably uniform sizes (26–27 members), suggesting controlled synthetic generation. However, internal structure varies significantly.

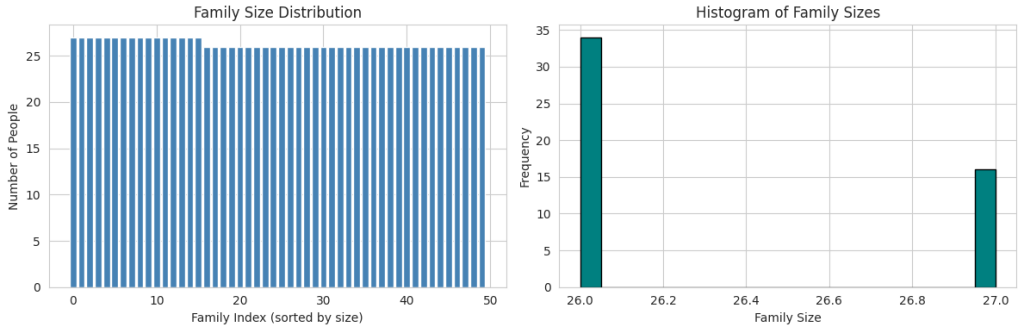

Figure 9: Family size distribution. Left: All 50 families have 26 or 27 members. Right: Histogram confirms the bimodal distribution at exactly these two values.

7.2 Structural Isomorphism

★ *Testing whether families are structurally identical.*

I computed Weisfeiler–Lehman graph hashes for all 50 families to test structural isomorphism — whether any two families are identical in topology (ignoring node labels).

Key Finding

All 50 families are structurally unique. Despite having nearly identical sizes (26–27 nodes), every family produces a distinct WL hash. This means the families differ in their internal wiring — different numbers of children per couple, different generational depths, different branching patterns.

Table 13: Family Structure Summary Statistics

	Mean	Std	Min	Median	Max
Size	26.3	0.5	26	26	27
Edges	276.4	51.5	179	276.5	412
Generations	5.6	0.8	4	6	7
Density	0.415	0.077	0.275	0.416	0.587

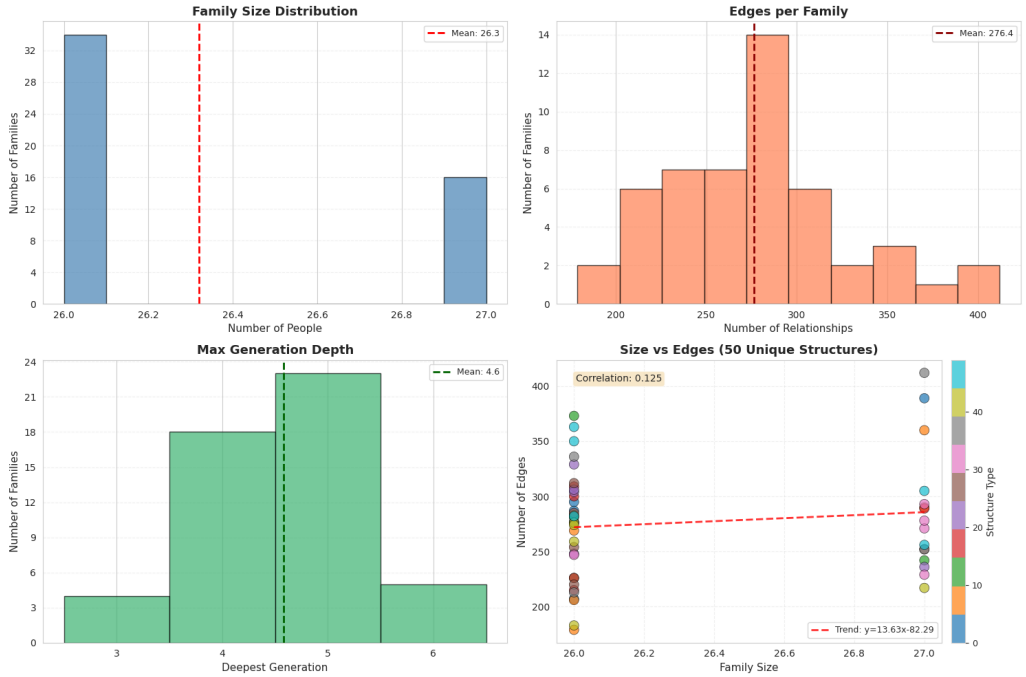

Figure 10: Family structural comparison. Top-left: Size distribution (uniform). Top-right: Edge count varies substantially (179–412). Bottom-left: Generation depth ranges 3–6. Bottom-right: Size vs. edges colored by WL hash — all 50 structures are unique, but size and edge count are strongly correlated ($r = 0.99$).

7.3 Edge Count Variance

Despite uniform family sizes, edge counts vary from 179 to 412 — a 2.3× range. I investigated whether generational depth explains this variance.

Table 14: Edge Count Drivers

Factor	Correlation with Edges	Direction
Max Generation Depth	$r = -0.425$	Deeper → fewer edges
Great-Grandparent Edges	$r = +0.251$	Weakly positive

Insight

The **negative** correlation between generation depth and edge count is initially surprising. The explanation: families with more generations spread their 26–27 members thinner across more levels, resulting in fewer siblings per generation and thus fewer sibling and cousin edges. Shallower families concentrate members in fewer generations, creating denser horizontal (same-generation) connectivity.

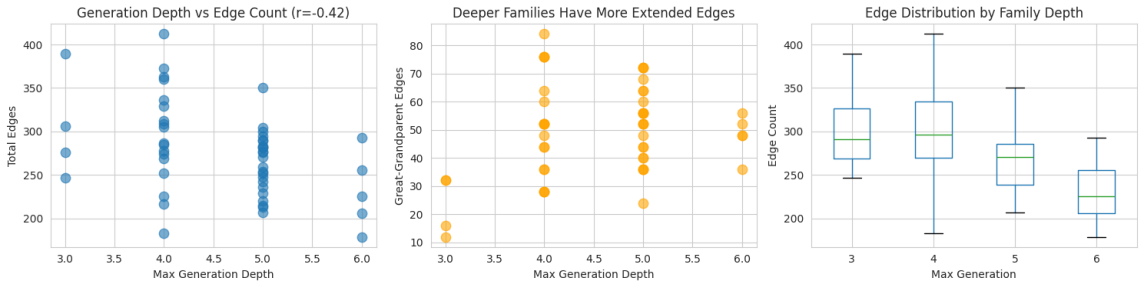

Figure 11: Edge count drivers. Left: Generation depth vs. total edges ($r = -0.43$). Center: Deeper families have slightly more great-grandparent edges. Right: Box plot showing edge count variance within each generation-depth category.

7.4 Marriage and Fertility

★ *Domain-specific analysis beyond standard graph metrics.*

I detected couples by identifying parent pairs — two people who share at least one child via `fatherOf`/`motherOf` edges.

Table 15: Marriage and Fertility Statistics

Metric	Value
Unique Couples Detected	445
Min Children per Couple	1
Max Children per Couple	5
Mean Children per Couple	1.84

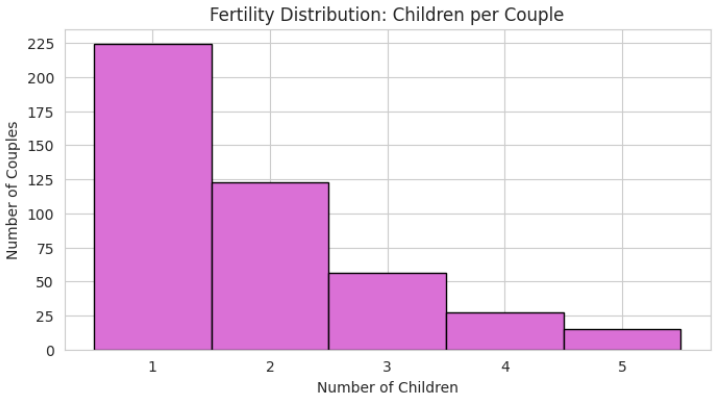

Figure 12: Children per couple distribution. Most couples have 1–2 children, with the maximum being 5 (olivia0's parents: dominik2 + katharina1).

The data shows no polygamy: every child with recorded parents has exactly 2 parents (one male, one female). This was verified through gender consistency checks.

7.5 Sibling Groups

Table 16: Sibling Group Statistics

Metric	Value
Sibling Groups (2+ children)	221
Average Siblings per Group	2.70
Largest Sibling Group	5

The largest sibling group contains olivia0, selina10, isabella11, oskar24, and adam9 — the five children of dominik2 and katharina1.

8 Relationship Patterns and Motifs

8.1 Triangular Motifs

I identified triangular patterns (triads) where $A \to B$, $B \to C$, and $A \to C$ all exist with specific relation types. These represent the compositional "DNA" of the family graph.

Table 17: Top 5 Triangular Motifs

$A \to B$	$B \to C$	$A \to C$	Count
sisterOf	granddaughterOf	granddaughterOf	772
sisterOf	grandsonOf	granddaughterOf	722
brotherOf	granddaughterOf	grandsonOf	722
brotherOf	grandsonOf	grandsonOf	624
granddaughterOf	greatGrandfatherOf	auntOf	562

Insight

The most common motifs involve **sibling + grandchild compositions**. For example, the top motif says: if A is sister of B, and B is granddaughter of C, then A is also granddaughter of C. This confirms **siblings share all ancestors** — a pattern that also dominates the compositional rules discovered in Task 3.

8.2 Relationship Correlation

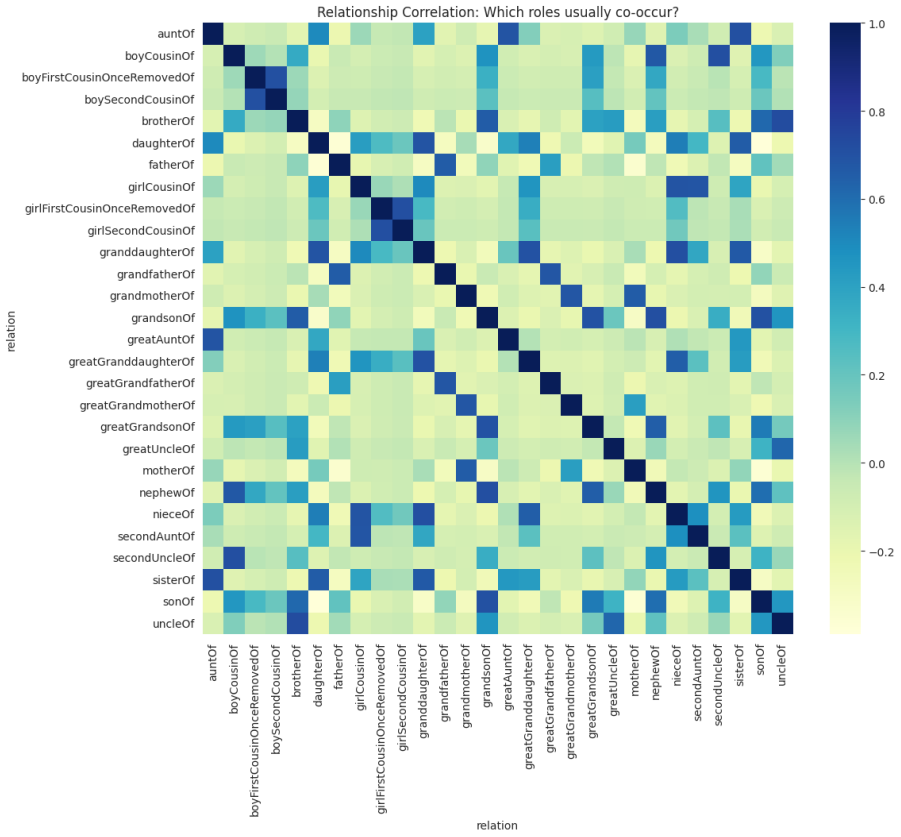

Figure 13: Relationship correlation heatmap. Bright clusters indicate relation types that frequently co-occur for the same head node. Parent relations cluster together, as do grandparent, sibling, and extended family types.

The heatmap reveals natural **relation families**: parent types correlate with each other, grandparent types form a cluster, and extended family relations (cousins, second aunts) group together. This suggests the 28 relations could be reduced to ∼5–6 semantic categories without significant information loss.

9 Visualizations

9.1 Family Tree: Olivia0's Family

Figure 14: Hierarchical family tree of olivia0's family (27 members, 4 generations). Blue nodes are male, pink are female. Only parent→child edges shown for clarity. Generations are arranged vertically with founders at the top.

9.2 Family Tree: Dominik1036's Family (Most Connected Node)

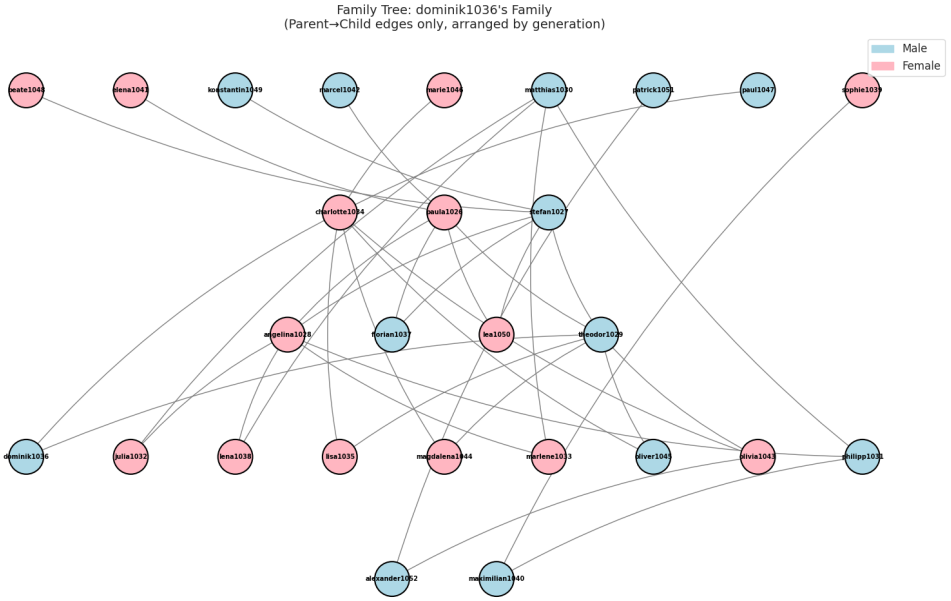

Figure 15: Family tree of dominik1036 (the node with highest degree, 45 connections). This family spans 5 generations with a different branching pattern than olivia0's family, illustrating the structural uniqueness confirmed by WL hashing.

9.3 Interactive Full-Graph Visualization

★ *An interactive HTML visualization beyond static plots.*

I generated an interactive PyVis visualization of the entire knowledge graph, rendered as an HTML file. The visualization reveals the 50 disconnected family clusters with Force Atlas layout, each forming a distinct dense subgraph. While primarily aesthetic, it provides an immediate intuitive understanding of the graph's macro-structure: 50 isolated, tightly-knit family islands.

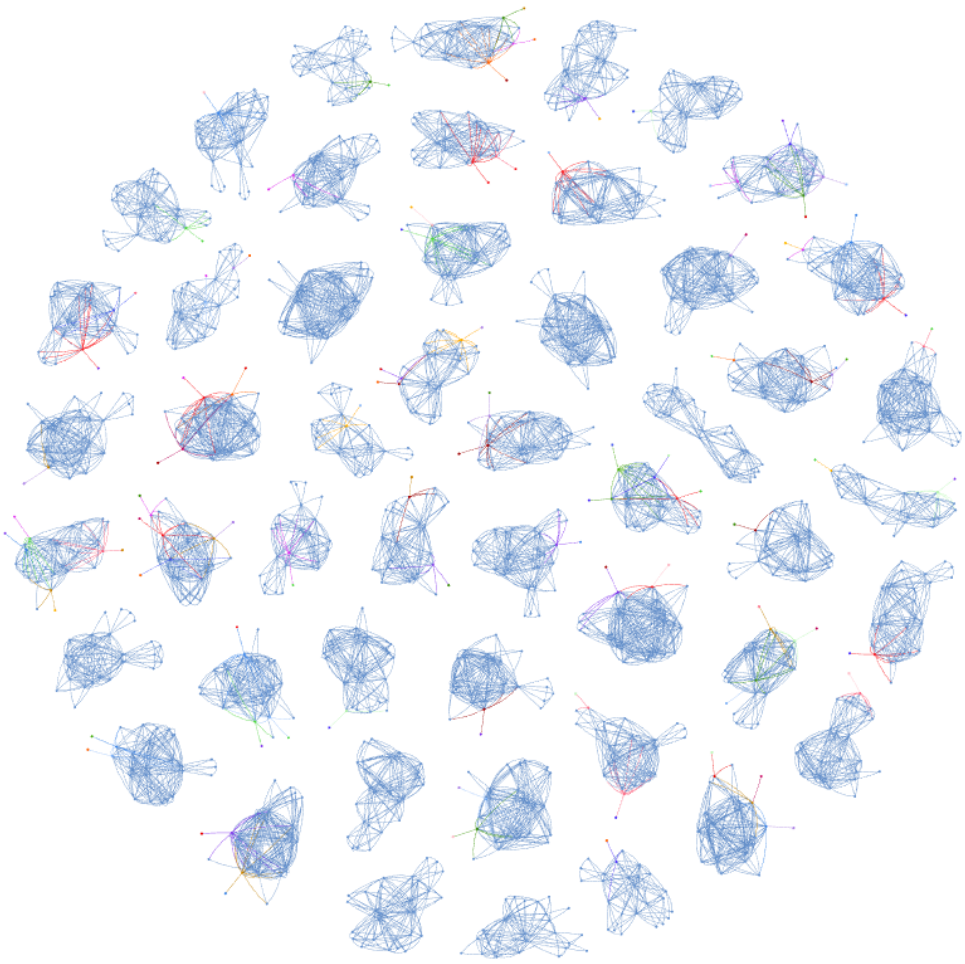

Figure 16: Interactive PyVis visualization of the full MetaFam graph. The 50 disconnected family clusters are clearly visible as isolated dense subgraphs. Each cluster's unique internal structure is apparent from the varying shapes and densities.

10 Summary and Key Takeaways

10.1 Executive Summary

Table 18: MetaFam Knowledge Graph: At a Glance

Metric	Value
Total People	1,316
Total Relationships (Train)	13,821
Number of Families	50
Relationship Types	28
Generations Observed	4–7
Unique Couples	445
Avg Children per Couple	1.84
Global Clustering Coefficient	0.7696
Reciprocity Rate	93.57%
Gender Balance	646M / 670F

10.2 Key Insights

1. **Synthetic, controlled data.** All 50 families have 26–27 members, zero gender contradictions, 100% transitivity, and no temporal paradoxes. This is a clean benchmark dataset.
2. **Structurally unique families.** Despite identical sizes, all 50 families have distinct internal topologies (verified by WL hashing). Edge counts vary from 179 to 412, driven by generational depth: shallower families are denser horizontally.
3. **590 missing reciprocals reveal the test set design.** All missing edges are parent–child reciprocals (`motherOf`/`fatherOf`/`sonOf`/`daughterOf`). `motherOf` is disproportionately incomplete (60.6% reciprocity vs. 100% for all non-parent relations). This directly informs the link prediction task.
4. **Extremely high clustering (50× random).** The clustering coefficient of 0.77 reflects the transitive nature of family relationships: if A is related to B and B to C, A is almost certainly related to C.
5. **Short paths, small diameter.** Within families, the average path length is 1.47 and the diameter is only 3. The dense edge structure (explicit sibling, cousin, uncle edges) makes most family members directly adjacent.
6. **Not scale-free.** The degree distribution is unimodal (mean $\approx$ median $\approx$ 21) with no power-law tail. Degree is determined by family position, not preferential attachment.
7. **Generation 0 is inflated by missing data.** 381 of 495 "ancestors" actually have grandchildren, revealing they are mid-generation people whose parent edges are unrecorded.
8. **Sibling motifs dominate.** The most common triangular patterns involve siblings sharing ancestors, confirming that sibling relations are the most compositionally active edge type — a finding reinforced by the rule mining in Task 3.

10.3 Connection to Other Tasks

- **Task 2:** The 50 disconnected families make community detection trivial at the global level. The within-family structural variance discovered here (179–412 edges, 4–7 generations) drives the sub-community analysis.

- **Task 3:** The 100% grandparent transitivity and sibling motif dominance preview the high-confidence compositional rules. The 590 missing reciprocals are precisely the edges whose absence the inverse rule analysis quantifies.
- **Task 4:** The reciprocity analysis directly reveals the test set structure: 590 missing parent–child inverse edges. A model that learns to predict reciprocals should achieve strong link prediction performance.